TOPOLOGY AND DYNAMICS OF COMPLEX SOCIAL NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Erik McCullough Volz

May 2006

TOPOLOGY AND DYNAMICS OF COMPLEX SOCIAL NETWORKS

Erik McCullough Volz, Ph.D.

Cornell University 2006

The problem of modeling complex social networks is considered from three perspectives: The problem of describing network topology; the problem of modeling dynamic processes on networks; and the problem of network sampling. These perspectives are highly complementary, each providing results with applications to one other. With respect to network topology, two main results are presented: An algorithm is presented capable of combining two measures of network structure, the degree distribution and the clustering coefficient. It is found that just two mechanisms are required to achieve any desired combination of these metrics– network growth, combined with preferential attachment. Secondly, a mathematical model of one class of complex network, semi-random networks, is presented which is capable of elucidating the structure of semi-random networks in greater detail then had been achieved with previous models. Among other results, this theory allows one to calculate the expected number of neighbors at a given distance from a randomly chosen node, and to compute the mean path length inside the giant component. Network dynamics are investigated with a simple epidemic model, the SIR (Susceptible Infected Removed) model. A mathematical theory is presented for predicting epidemic incidence for SIR dynamics in semi-random networks. Finally, the problem of network sampling is considered. A probability based estimation theory is presented for Respondent Driven Sampling (RDS). The

theory enhances RDS by offering greater analytical tractability, analytical variance estimation, and the estimation of means of continuous variables.

BIOGRAPHICAL SKETCH

Erik Volz received his BA in Mathematics and Russian language from the University of Rochester in 2002. In the Fall of 2002, he began PhD studies in the Department of Sociology, Cornell University. Erik received his MA in sociology in Spring of 2004 with distinction.

Erik's life has followed a boring trajectory consistent with the life-histories of successful academics. The only major events worth reporting concern standard academic milestones. For this, he apologizes to the reader.

This work is dedicated to my father, who instilled in me a love of science and knowledge.

ACKNOWLEDGEMENTS

Thanks to Douglas Heckathorn and Matt Salganik for valuable suggestions concerning Respondent Driven Sampling. Steve Strogatz has been a tremendous moral support and has given valuable suggestions for adapting this work for diverse interdisciplinary audiences. Victor Nee has given valuable support during the manuscript's completion. Steve Ellner offered valuable advice and references during the early stages of the research. My Chair, Doug Heckathorn has been an excellent mentor and guide for research occupying an interdisciplinary space on the fringe of sociology.

Thanks especially to Cornell University, the United States government, and the American taxpayer. Together, these entities have created a brilliant and equitable research institution, where any individual with sufficient motivation can pursue intellectual fulfillment.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vi
List of Tables . . . viii
List of Figures . . . ix

1 Introduction 1

2 Random networks with tunable degree distribution and clustering 7
2.1 Random network model . . . 9
2.2 Results . . . 17
2.3 Variations on the agorithm . . . 27
2.3.1 Methods for generating degree assortativity . . . 28
2.4 Methods for generating lists of potential triads . . . 29
2.5 Phase transitions . . . 30
2.6 Finite size effects . . . 33
2.7 Dependence of the clustering coefficient on input parameter C_{input} . . . 35
2.8 Implications for sociology . . . 37
2.9 Discussion . . . 39

3 Tomography of random social networks 45
3.1 Network tomography . . . 46
3.1.1 Descriptive statistics . . . 52
3.2 Theoretical Examples . . . 54
3.3 Email Network . . . 60
3.4 Discussion . . . 67

4 SIR dynamics in populations with heterogeneous connectivity 73
4.1 Intuitive model specification . . . 75
4.2 Formal model specification . . . 81
4.2.1 Definitions . . . 81
4.2.2 Dynamics . . . 83
4.3 Examples . . . 86
4.4 Discussion . . . 93

5 Probability based estimation theory for Respondent Driven Sampling 99
5.1 Respondent Driven Sampling . . . 100
5.2 New estimators for Respondent Driven Sampling . . . 107
5.3 The classical RDS estimation procedure and its relation to RDS II 111
5.4 Variance estimation . . . 115
5.5 Simulation study of RDS I and RDS II . . . 118

5.6 Discussion 124

LIST OF TABLES

2.1 Detailed description of the clustering method. 13
2.2 Detailed description of the clustering method continued. 14

4.1 A summary of the nonlinear differential equations used to the describe the spread of a simple SIR type epidemic through a random network. The degree distribution of the network is generated by $g(x)$. 74

5.1 Notation used throughout this paper. 104
5.2 Random networks were generated with four disjoint groups, each having the size N_X and Poisson degree distribution with average degree z. 120
5.3 RDS I, RDS I/DS, and RDS II are compared for a real data set. The data come from a survey of 264 New York city jazz musicians [7]. 124

LIST OF FIGURES

2.1 Overview of the network construction process. The first node (far left) is chosen at random. Then neighbors for that node are chosen as described in the text. Subsequently, neighbors are chosen for the new nodes, but now we have new connections formed with nodes two steps away with probability C_{input}. Triadic connections are indicated with dotted lines. This process continues until the waves die out, and a new component is formed, or all nodes are exhausted. 12

2.2 Two examples of networks generated with the algorithm. Left: Random network with power law degree distribution, $\kappa = 15$, $\gamma = 2$, $C = 0.15$. Right: Random network with poisson degree distribution, $z = 4$, $C = 0.40$. Note that these are abstract representations of random networks. The spatial embedding of the network does not have any meaning. 15

2.3 Random graphs were generated with an exponential degree distribution ($\lambda = 1.4$) with two algorithms: 1. The clustering algorithm described in this text with $C = 0$ 2. A "stub-matching" algorithm as in [28], known to produce true random graphs with specified degree distributions. The frequency of component sizes is illustrated above. 18

2.4 Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.00. Compare with figures 2.2(right) and 2.8. 19

2.5 Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.30 20

2.6 Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.40. 21

2.7 Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.60. The image is zoomed on several of the largest components. 22

2.8 Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.97 23

2.9 Size of the giant component versus the clustering coefficient in a poisson random network, $z = 3$. Each point represents the average of 40 trials. 24

2.10 N=5,000 nodes. Power law with parameters $\kappa = 10$ and $\gamma = 2$. Each point represents the average of 40 trials. Compare this with 2.9. The phase transition is much less sharp than for the poisson random networks. 25

2.11 Two random networks are compared over a range of parameter values for the power law degree distribution with parameters κ and $\gamma = 2$. Each point represents the average of 40 trials. 26

2.12 The size of the giant component is shown versus the input clustering parameter C_{input}. The network is Exponential(4), $n = 20000$ . . . 28

2.13 The size of the giant component is shown vs. z, the parameter of the poisson degree distribution, for four levels of clustering ($C = 0.0, C = 0.15, C = 0.30, C = 0.40$). The vertical lines indicate the point of the phase transition for each level of clustering predicted by equation 2.4 33
2.14 The percentage reduction in the number of "stubs" is shown versus the Clustering Coefficient for two networks: (i) Poisson degree distribution with parameter = 4, (ii) Exponential degree distribution with parameter = 2. N=5000 for both networks. Each point is based on the average of 20 trials. 34
2.15 The percentage reduction in the number of stubs is shown versus the network size. The network has a Poisson degree distribution with parameter = 4, C = 0.80. Each point is based on the average of 20 trial networks. 36
2.16 The clustering realized versus the input clustering parameter C_{input}. The random network has a poisson degree distribution with $z = 8$. $N = 2500$. 37

3.1 This diagram illustrates the tomographic method detailed in the text. Starting from a single node v_0 we recursively explore nodes at distance l from v_0. R_l is the number of connections going to layer l from layer $l-1$. S_l is the number of connections to nodes in layer l. T_l is the number of connections not connected to nodes in layer l or less. The importance of these quantities is explained in the text. 48
3.2 $n = 50000$, Poisson degree distribution, $z = 3$. Data points are the average of 40 generated networks with 20 trials per network. Solid lines represent the theoretical prediction given by 3.17. 57
3.3 $n = 50000$, Poisson degree distribution, $z = 1.25, 3, 5$. Data points show the 10'th and 90'th percentile for 40 randomly generated networks with 20 trials per network. Solid lines represent the theoretical prediction given by 3.18. 58
3.4 $n = 50000$, Exponential degree distribution, $z = 3$. Data points are the average of 40 generated networks with 20 trials per network. Solid lines represent the theoretical prediction given by 3.11. 59
3.5 $n = 50000$, Exponential degree distribution, $z = 1.25, 3, 5$. Data points show the 10'th and 90'th percentile for 40 randomly generated networks with 20 trials per network. Solid lines represent the theoretical prediction given by 3.7. 61

3.6 The giant component from the Cornell email network. Connections in the network represent reciprocal communication within a 24 hour sampling frame. The nodes are color-coded. Blue nodes are faculty, red nodes are graduate students, green nodes are undergraduates, and yellow nodes are everyone else, mainly administrators. The network 2607 nodes and 4838 connections. The giant component consists of 1227 nodes. 62
3.7 Degree distributions for the reciprocal and non-reciprocal email networks. Solid lines show a fit designed to match the average degree of the empirical distribution. The theoretical density is given by equation (3.19). 63
3.8 Theoretical (solid line) and empirical (dotted line) stratum sizes for the R/NR email network. This network includes both reciprocal and non-reciprocal communication within the 24 hour sampling frame. The upper dotted line represents 90'th percentile stratum sizes picking a *seed* from the network uniformly at random. The lower dotted line represents the 10'th percentile. 65
3.9 Theoretical (solid line) and empirical (dotted line) stratum sizes for the R email network. This network includes only reciprocal communication within the 24 hour sampling frame. The dotted line represents the mean empirical stratum size, selecting a *seed* from the network uniformly at random. 66

4.1 The number of infecteds (including recovered) is shown versus time for an SIR model on three networks. Force of infection and mortality are constant: $r = 0.2$, $\mu = 0.1$. The networks have Poisson ($z = 3$), power law ($\gamma = 1.615, \kappa = 20$), and exponential ($\lambda = 3.475$) degree distributions. Each of these degree distributions has an average degree of 3. 88
4.2 α, β, and W/n are shown versus t for a power law network with exponent $\kappa = 1.615$ and exponential cutoff $\kappa = 20$. Force of infection and mortality are constant: $r = 0.2$, $\mu = 0.1$. 89
4.3 The degree distribution (equation (4.32)) for susceptibles is shown at three different times during the course of an epidemic on a Poisson network ($z = 3$). Force of infection and mortality are constant: $r = 0.2$, $\mu = 0.1$. 90
4.4 The number of infecteds (including recovered) is shown versus time for an SIR model on a Poisson network ($z = 3$). Each of these trials are below the critical level of transmissibility required to sustain an epidemic. Mortality is constant, $\mu = 0.4$, while three different levels of the force of infection are tried, $r = 0.15, 0.17, 0.18$. 92

5.1 Example of a recruitment chain. This recruitment chain comes from a RDS study of jazz musicians in New York City [7]. Arrows indicate the direction of recruitment. The colors indicate the gender of each respondent: Black = Male, White = Female, Grey = Missing Data 102
5.2 Variance of three RDS estimators and mean estimated variance, based on 50,000 simulations as described in the text. Sample size is varied from 75 to 500. The data are plotted with log-log axes. . . 121
5.3 Variance of RDS II and RDS I, alongside the estimated variance for RDS II based on 50,000 simulations with sample size 500 as described in the text. The mixing parameter σ_{AA} is varied from 0.069 to 0.57. 123

CHAPTER 1

INTRODUCTION

The patterns of human contact and interaction have long been of interest to sociologists [14, 7, 15, 10], psychologists [11, 13], epidemiologists [8, 2, 3, 5, 6], and lately numerous scientists from the mathematical sciences [4, 9, 1, 16, 12]. Developing accurate descriptions of social networks is of both scientific and practical importance–scientific, because those who study social networks find structures of deep beauty and intricacy, while increasing our understanding of contact patterns holds the promise to unlock a deeper understanding of fundamental sociological processes. Social networks are also relevant in matters of public import such as the spread of infectious diseases.

The most accurate description of human contact patterns is the social network– a combinatorial device which describes in a binary way whether two given individuals are connected to one another or interacting. The network model has recently grabbed the attention of population modelers. This dissertation is exclusively concerned with network models of populations. It has been completed on the backdrop of rapid advances in network theory over the last decade. The contemporary study of complex networks was kicked off less than 10 years ago with the study of a social network problem– the small-world effect [13], which is the empirical observation that highly clustered social networks tend to also have very short mean-path length. In simple terms, this means that there are relatively few intermediaries connecting any two members of a social network, and this occurs despite the fact that a randomly chosen person is likely to already know most of the friends of his friends. Mathematicians [15, 16] have been very successful in modeling this effect, and have brought the problem of the effects and causes of the small-world

phenomenon to a satisfactory conclusion. On the heels of these results, researchers began to notice another puzzling feature of many social networks, the apparently scale-free, power-law distribution in the number of contacts to and from each individual. Similar to the small-world problem, applied mathematicians [1] have been quick to offer mechanisms which explain this phenomenon. These two problems have formed the backdrop on which subsequent complex networks research has developed. There has been a great deal of progress made in understanding the various statistics which characterize the structure of complex networks, such as the degree distribution, clustering coefficient, and mean path length. The present work makes further contributions in this regard.

As with any attempt to study a complex natural phenomenon, one encounters a problem when investigating complex social networks– namely that the mechanisms which are responsible for the observed structure work together unpredictable and nonlinear ways. As soon as one has come to a partial understanding of one factor shaping social networks, other factors appear which warrant consideration. The challenge of modeling complex social networks seems to have irreducible depth. Ultimately the most accurate description of a social network is the network itself. Yet we can gain insight into this structure by focusing on the fundamental mechanisms responsible for producing this structure. Most of all, this requires patience. It is unwise to model networks using a plethora of details and factors, as the resulting model will tend to be as incomprehensible as the empirical network. It is a far wiser approach to begin with simple models, and then to proceed only when one factor is thoroughly understood.

I am inclined to approach the study of complex networks with as much mathematical analysis as possible. Bringing mathematics to bear on a complex problem

necessitates a great deal of simplification and approximation. Sometimes my approximations and assumptions will seem crude. Yet this is merely an extension of the patient approach to understanding complex networks. It is most desirable to understand a simple model with mathematical precision before the next most complex problem is taken on.

The papers presented here represent the culmination of four years of investigation into the problem of understanding complex networks. The topics I address run the gamut of complex networks research, from topology, to dynamics, and finally the problem of network sampling. Understanding each problem individually contributes to a greater understanding of the whole. Network sampling can be used to gain insight into the structure of real social networks, which can then be modeled using our theory of network topology. Once network topology is understood, the ultimate aim of understanding network dynamics is within reach.

The methods I employ, mathematical and computational, should be of interest to researchers studying complex networks in a variety of disciplines. Yet when exploring empirical applications, it is my desire to focus on social networks– those networks which have importance to us in our daily lives, affecting everything from our ability to find a job, to the likelihood we will catch a flu. But the results given here could find equal application for those studying technological networks such as the Internet or biological networks such as food webs.

The first two chapters focus on problems of complex network topology. First I consider the mechanisms which allow real networks to organically combine metrics such as the degree distribution and clustering coefficient in ways which are difficult to reproduce from a top-down design perspective. Secondly, I develop a mathematical theory for elucidating the structure of the simplest class of com-

plex networks, semi-random nets, which are random with respect to a specified degree distribution. The approach, dubbed "network tomography", analyzes the structure of semi-random networks from the ego-centric perspective of a randomly chosen node.

The third chapter applies the methods developed on the tomography problem to the problem of modeling epidemics on networks. The signature feature of many real-world populations is the heterogeneity in the number of contacts a given individual has. Each contact presents a possible avenue for infection. Taking this heterogeneity into account has been a persistent challenge for mathematical epidemiologists. Yet it is possible to use a similar approach to that taken in the tomography paper to take this heterogeneity into account. The primary result of this chapter is a system of three differential equations to describe the epidemic incidence for SIR type contagion in populations with arbitrary degree distributions.

The final chapter considers the problem of network sampling. This research builds on Respondent Driven Sampling, a chain-referral method which harnesses the social network of the target population to collect a analytically tractable sample. Combined with a mathematical model of the sampling process, it is possible to make unbiased estimates of the target population. My research builds on previous research by simplifying RDS estimation theory, yielding a more tractable probability-theoretic approach. Other enhancements include the ability to estimate means of continuous variables and analytical variance estimation.

REFERENCES

[1] R. Albert and A.L. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47, 2002.

[2] L. Ancel-Meyers and M.E.J. Newman. Applying network theory to epidemics: control measures for mycoplasma pneumoniae outbreaks. *Emerging Inf. Dis.*, 9:204, 2003.

[3] T.D. Eames and M.J. Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *PNAS*, 99:13330–13335, 2002.

[4] P. Erdős and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[5] S. Eubank, H. Guclu, V.S. Anil-Kunar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic social networks. *Nature*, 429:180–184, 2005.

[6] S. Gupta, R.M. Anderson, and R.M. May. Networks of sexual contacts: Implications for the pattern of spread of hiv. *AIDS*, 3:807–817, 1989.

[7] J.Scott. *Social Network Analysis: A Handbook*. Sage, London, 2nd edition, 2000.

[8] M. Kretzschmar and M. Morris. Measures of concurrency in networks and the spread of infectious diseases. *Math. Biosciences*, 133:165–195, 1996.

[9] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, 2001.

[10] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99:2566–2572, 2002.

[11] A. Rapoport. *Handbook of Mathematical Psychology*, chapter Mathematical models of social interaction. Wiley, New York, 1963.

[12] S.N.Dorogovtsev and J.F.F. Mendes. *The evolution of networks: from biological nets to the Internet and WWW*. Oxford University Press, Oxford, 2003.

[13] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425, 1969.

[14] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, Cambridge, 1994.

[15] D.J. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, 1999.

[16] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

CHAPTER 2

RANDOM NETWORKS WITH TUNABLE DEGREE DISTRIBUTION AND CLUSTERING

Many random network models have been proposed to replicate important aspects of the topology of real-world networks [27, 28, 1, 2, 5, 29, 7, 36, 34, 26, 3, 18, 32, 19]. In particular, much attention has been paid to the degree distribution and the clustering coefficient. A great deal of progress has been made on network models which combine certain degree distributions with some level of clustering [9, 17, 32, 10, 16, 13]. It has been an open problem to combine these two topologies in the most general way. Is it possible to have a network model which is flexible enough to accommodate any combination of degree distribution and clustering? In this article we propose such a model and demonstrate its effectiveness by generating networks over a wide range of parameters.

Random network models have fallen in several broad categories. Some models have focused on Monte Carlo techniques to reproduce a specific topology [27, 28, 22]. Other models have specific topologies built into them (e.g. regular lattices) in order to explicate the so-called small-world problem [36, 34]. Yet other models have focused on plausible mechanisms for how networks form, such as a growth process with preferential attachment [9, 26, 3]. In common with most mechanism-based models, we produce our networks by growing them from one initial node. We find that being able to construct a network one node at a time also offers sufficient flexibility to combine arbitrary degree distributions and clustering.

Once we have a network model which can combine arbitrary degree distributions and clustering, it is of interest to explore the effects of these parameters on the size of the giant component and the point of the phase transition where a

giant component forms. This is true with regard to clustering in particular, as so far models capable of interpolating between extremes of this parameter have been lacking. In section 2.2 we explore the effects of clustering on the size of the giant component and point of the phase transition. In section 2.5 we present some analysis.

Throughout this article we will rely on the following definitions: The *degree distribution* of a network describes how many neighbors a node in a network has. The probability of a node having degree k in a network is described by the degree distribution p_k, where p_k can take the form of any well defined discrete density function over the positive integers. Examples frequently employed in the literature are

- Poisson: $p_k = \frac{z^k e^{-z}}{k!}, k \geq 0$
- Power-law. For our experiments, we use power-laws with finite cuttoffs κ: $p_k = \frac{k^{-\gamma} e^{-k/\kappa}}{Li_\gamma(e^{-1/\kappa})}, k \geq 1$ where $Li_n(x)$ is the nth polylogarithm of x.
- Exponential: $p_k = (1 - e^{-1/\lambda})e^{-k/\lambda}, k \geq 0$
- Empirical: The degree distribution is estimated from a network sample.
- Gaussian: The ordinary Gaussian must be modified to be positive and discrete.

The *clustering coefficient* C describes the proportion of triads in a network out of the total number of possible triads. The clustering coefficient is defined:

$$C = \frac{3N_\Delta}{N_3}$$

where N_Δ is the number of triads in the network and N_3 is the number of connected triples of nodes. Note that in every triad there are three connected triples.

There is also a measure of *local Clustering* given by

$$C_i = \frac{N_\Delta(i)}{\binom{\delta(i)}{2}}$$

where $N_\Delta(k)$ is the number of triads connected to node i, $\delta(i)$ is the degree of node i, and $\binom{\delta(i)}{2}$ is the number of potential triads connected to a node of degree $\delta(i)$. The average value of local clustering (i.e. "Watts-Strogatz Clustering" [36]) is also of interest:

$$\frac{\sum C_i}{N}$$

where N is the number of nodes in the network. This value is frequently close to the clustering coefficient, and will be equal to the clustering coefficient if local clustering is constant throughout the network.

2.1 Random network model

Introducing clustering into a network with a specified degree distribution is a non-trivial problem. Any method aspiring to introduce an arbitrary amount of clustering into a network must interpolate between two extremely different topologies. When clustering is 0%, the method must reproduce pure random networks with specified degree distributions. When clustering is 100%, there is only one configuration a network may have: each node must be connected to a small clique where every node has the same degree, and all of a node's neighbors are connected with one another. This challenge is made all the more difficult by trying to make the model networks general enough to accommodate any desired degree distribution.

The most obvious way of introducing triads is to simply define a *rewiring rule* whereby links are swapped between nodes so as to introduce triads while leaving the degree distribution the same. Such rewiring schemes quickly run into problems, as

it is impossible to define a rule such that the number of triads is strictly increasing and the number of triads introduced does not max out. The problem is that when links are "swapped" among nodes, triads are not only created but can be destroyed. For example, in our simulations we have found that such schemes are effective only for introducing about 15% clustering into a poisson random network.

Rewiring algorithms have proven effective at the related challenge of adjusting the *average local clustering*. Kim [18] has recently used rewiring algorithms to introduce large amounts of *local clustering* into networks. Using a MC simulations at zero-temperature (i.e. a triad is never destroyed in the rewiring process) and a Hamiltonian of $\sum -C_k$, Kim was able to modify various networks with diverse degree distributions to exhibit average local clustering ($\sum C_k/N$) ranging from 0% to 70%.

Newman [25] and Guillaume et al. [13] have had some success with another approach. These authors define a bipartite network of individuals and affiliations. Then they project the bipartite network onto a unipartite network of only nodes and no affiliations by connecting two nodes if they share a common affiliation. The distributions of affiliation size and the affiliation-degree distribution of the nodes is chosen in such a way as to produce a desired level of clustering. Tuning the degree distribution simultaneously has proven more challenging, however. While the bipartite projection method may actually have the potential to generate pure random networks with tunable degree distributions and clustering, so far it's efficacy has only been shown for exponential and power-law random networks. It remains an open problem to implement it for arbitrary degree distributions.

Our method works by growing networks. The algorithm first initializes all nodes with a degree drawn i.i.d. from the desired degree distribution. Then the random

network is constructed by an iterative procedure similar to a branching process. The premise is to start from a single node and then assign new connections entirely at random under the constraint that a certain amount of clustering must exist. The algorithm is described in detail in table 2.1, and is schematized in figure 2.1. Two example networks are shown in figure 2.2.

Our model has similarities and differences with other models proposed in the literature. Like the algorithm of Milo et al. [22], each node is assigned a unique degree prior to any edges being formed between nodes. But like the model networks of Barabasi [1], Dorogovtsev et al. [21] among others, the network is constructed via a growth process. The first node is chosen at random, and subsequently nodes are added to the graph by attaching them to nodes which still have stubs that have not been matched. When the new node forms its own connections, it first forms a list of all nodes which are two steps away. Then with probability C_{input}, that node is selected as the next neighbor.

One complicated feature of this algorithm concerns the probability of selecting a new neighbor from the stub list. In fact, new neighbors cannot be selected uniformly at random from the stub list, as clustering implies a certain amount of degree assortativity among the nodes in the network. For example, a node connected to a degree k node has $k-1$ potential triads in common with that node, and on average will have $C(k-1)$ common triads. This implies that the node must have on average a degree at least equal to $C(k-1)$.

Because triads are distributed uniformly throughout the network, the number of triads connected to a node of degree k is distributed $binomial(\binom{k}{2}, C)$. As noted above the number of common triads with a neighbor of degree k is distributed $binomial(k-1, C)$. Let τ_{ij} denote the number of triads node i has in common

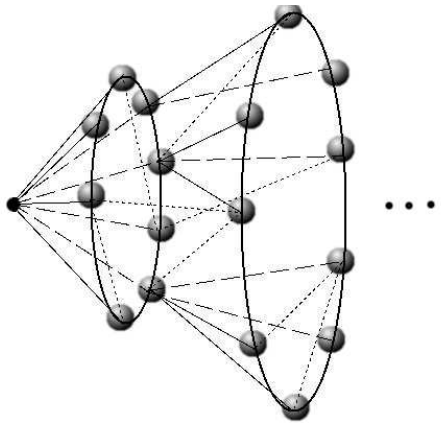


Figure 2.1: Overview of the network construction process. The first node (far left) is chosen at random. Then neighbors for that node are chosen as described in the text. Subsequently, neighbors are chosen for the new nodes, but now we have new connections formed with nodes two steps away with probability C_{input}. Triadic connections are indicated with dotted lines. This process continues until the waves die out, and a new component is formed, or all nodes are exhausted.

Table 2.1: Detailed description of the clustering method.

1. Initialize all nodes with a degree drawn i.i.d. from the degree distribution

2. Form a list of "stubs"– connections of nodes which have not yet been matched with neighbors. Call this list StubList.

3. Pick a starting node, v_0, uniformly at random from all nodes.

4. For each of v_0's stubs, choose a new neighbor by picking an element v_1 from the stublist with probability $p_{v_1|d(v_0)}$ as described in the text. If the new neighbor is not

   - the same node as v_0
   - already connected to v_0

   then form the connection. Otherwise, repeat the process until a valid neighbor is found. Add all of the new neighbors from this process to a list called NextWave.

5. Copy all elements of NextWave to a list called CurrentWave. Remove all elements from NextWave. For all elements in CurrentWave:

This is continued in table 2.2.

Table 2.2: Detailed description of the clustering method continued.

(a) Form a list of all nodes 2 steps away. If a node does not have any stubs left in StubList, throw it out. Call this list PotentialTriads

(b) For all stubs which have not been assigned neighbors

   i. Scan through PotentialTriads. With probability C^{input}, connect to node $v_3 \in$ StubList. Remove element v_3 from PotentialTriads regardless of whether it was selected. If it was selected, also remove an instance of v_3 from the StubList.

   ii. If no neighbors were selected from PotentialTriads, select a new neighbor by choosing from StubList as above. If the new neighbor is not in CurrentWave, and if the new neighbor is not already in NextWave, add them to NextWave.

Repeat the last step until NextWave is empty following an iteration. Then, if StubList is empty, the process is complete– all connections have been formed. Otherwise, start a new component by choosing a new starting node uniformly at random from those not yet in the network.

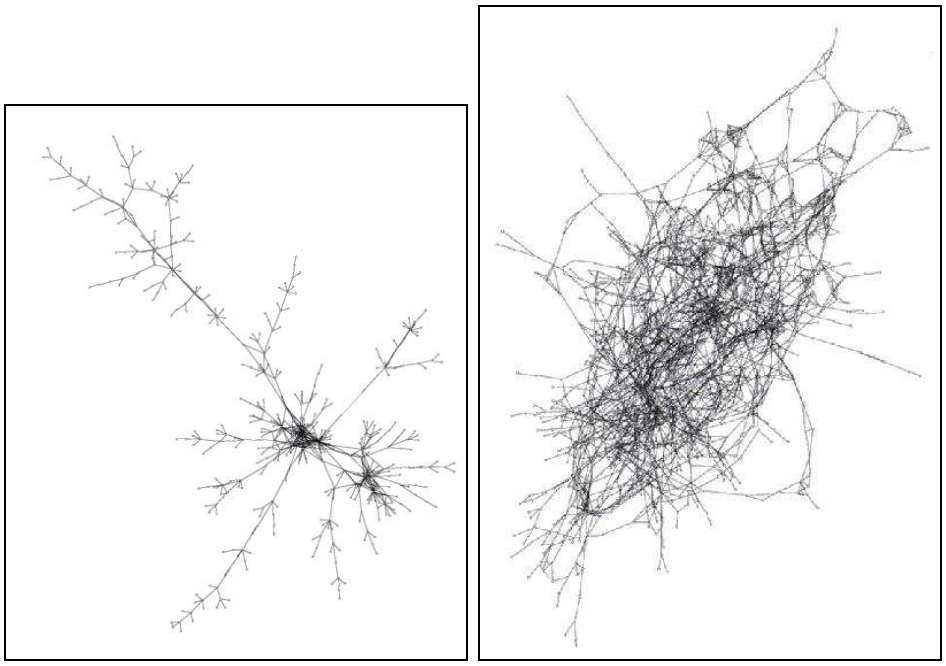
1. 

Figure 2.2: Two examples of networks generated with the algorithm. Left: Random network with power law degree distribution, $\kappa = 15$, $\gamma = 2$, $C = 0.15$. Right: Random network with poisson degree distribution, $z = 4$, $C = 0.40$. Note that these are abstract representations of random networks. The spatial embedding of the network does not have any meaning.

with node j, and τ_{ji} denote the number of triads j has in common with i. Of course these two random variables should be equal. We can calculate the probability of these two potential neighbors as having an equal number of common triads as:

$$p_{ij}^c = \sum_{x=0}^{min\{d(i),d(j)\}} p(\tau_{ij} = x)p(\tau_{ji} = x)$$

Let q_j denote the probability of selecting node j from the stub list. Then the correct probability for selecting node j as a neighbor is:

$$q_{ij} = \frac{q_j p_{ij}^c}{\sum_\alpha p_{i\alpha}^c}$$

which is just q_j weighted by the probability of the two neighbors having a compatible number of triads in common.

In order to sample from this distribution, we use Markov Chain Monte Carlo techniques. For a large number of iterations we select a new node β from the stub list, then with probability $a_{\alpha\beta}$ we accept this new neighbor, where α is the currently selected node in the markov process, and

$$a_{ij} = \frac{p_{i\mu}^c}{p_{i\alpha}^c}$$

If β is not accepted, we keep α for the next iteration. The final neighbor is the node selected at the last iteration.

It is desirable that our algorithm selects networks as uniformly as possible from the ensemble of all networks which realize a given degree distribution and clustering coefficient. It is difficult to prove that our algorithm is truly unbiased in this sense, though our networks do have many of the properties of an unbiased random network. The algorithm can be tuned to produce exactly the right proportion of triads to triples in the limit of large graph size. Furthermore, the degree of the nodes were chosen as i.i.d. random variables, so in the limit of large graph size, the

degree distribution is unbiased too. Triads are uniformly distributed throughout the network as reflected by the fact that the local clustering is independent of degree. Lastly, when this algorithm is used to produce networks with no clustering at all, it produces networks with the same statistical properties as true random graphs with a specified degree distribution. As shown in figure 2.3, the distribution of component sizes for networks made with this algorithm is identical to true random graphs with specified degree distribution without clustering.

It is worth noting that many real-world networks, particularly in the biological realm, have local clustering which scales as $1/k$ [30]. Our model in contrast produces constant local clustering, though it may be possible to generalize our method to create networks with any desired schedule of local clustering.

2.2 Results

We have explored the effects of clustering and degree distribution over a wide range of parameters. Figures 2.2(right), 2.4, and 2.8 illustrate the effect of clustering on the structure of a random networks with poisson degree distributions ($z = 3$) as clustering is increased from 0 to 1.00[1]. As C is increased, nodes tend to disaggregate into smaller tightly connected clusters of nodes with similar degree. This has the overall effect of decreasing the giant component size as clustering is increased. In the limit as C goes to 1, we find that the network breaks down into many small completely connected cliques with each node in a clique sharing a common degree.

Figure 2.9 shows the effects of clustering on the size of the giant component for a

[1]All networks were rendered with yEd © http://www.yworks.com, free for non-commercial use.

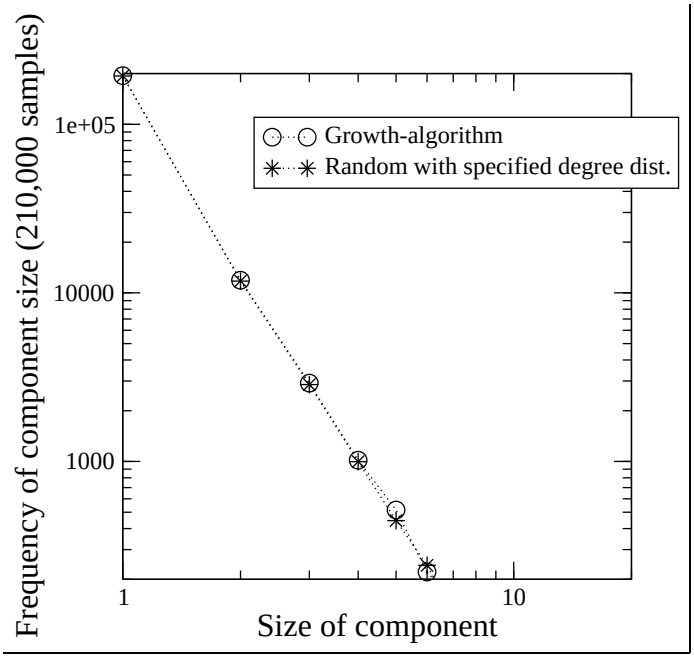


Figure 2.3: Random graphs were generated with an exponential degree distribution ($\lambda = 1.4$) with two algorithms: 1. The clustering algorithm described in this text with $C = 0$ 2. A "stub-matching" algorithm as in [28], known to produce true random graphs with specified degree distributions. The frequency of component sizes is illustrated above.

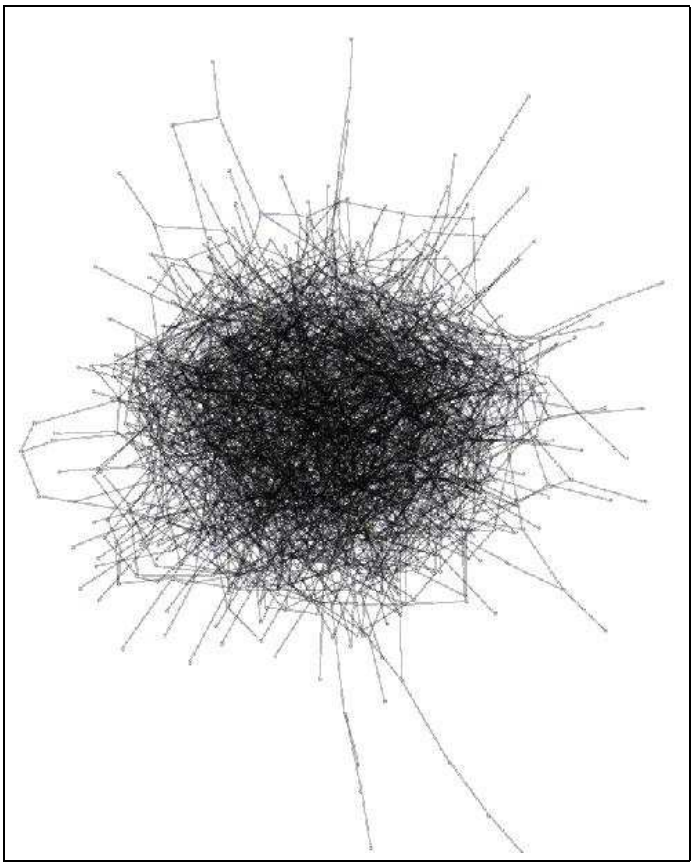


Figure 2.4: Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.00. Compare with figures 2.2(right) and 2.8.

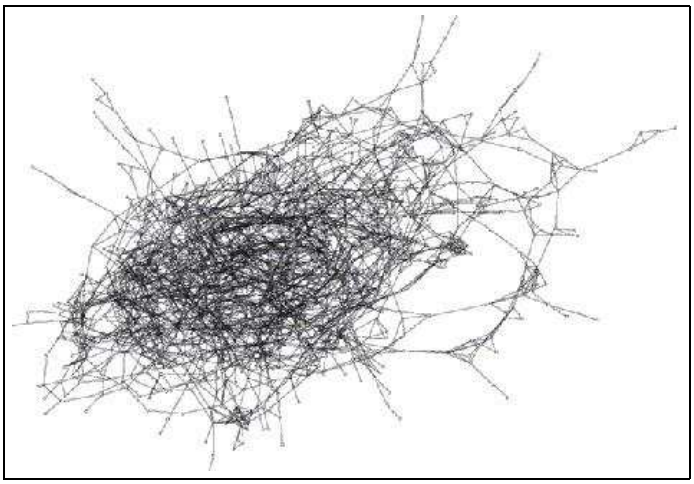


Figure 2.5: Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.30

poisson random network. Clustering varies from 0.05 to 0.90. The giant component seems to undergo a phase transition at a critical level of clustering around $C = 0.60$. In the next section we will find that the critical clustering value is actually $C^* = 0.618$. At this point, nodes suddenly disaggregate into much smaller, tightly inter-connected groups. Similar phase transitions have been observed throughout the networks literature, particularly concerning the targeted deletion of links and nodes in percolation phenomena [33]. This algorithm has similar disconnecting results without modifying the degree distribution of the network.

Regarding power-law networks (see figure 2.10), we note the striking tendency for moderate levels of clustering to limit the size of the giant component. Because the number of potential triads connected to a node scales as k^2, the high degree vertices account for most of the clustering. In networks with highly skewed degree distributions, the high-degree nodes must connect to one another in order to realize the required number of triads. This has the effect of limiting the ability to act

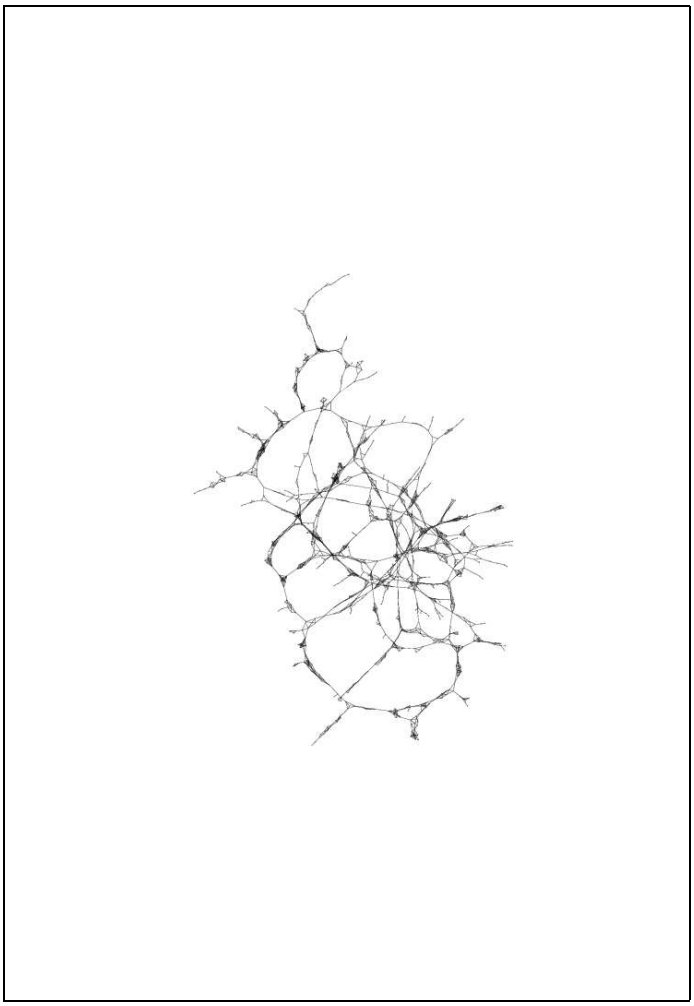


Figure 2.6: Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.40.

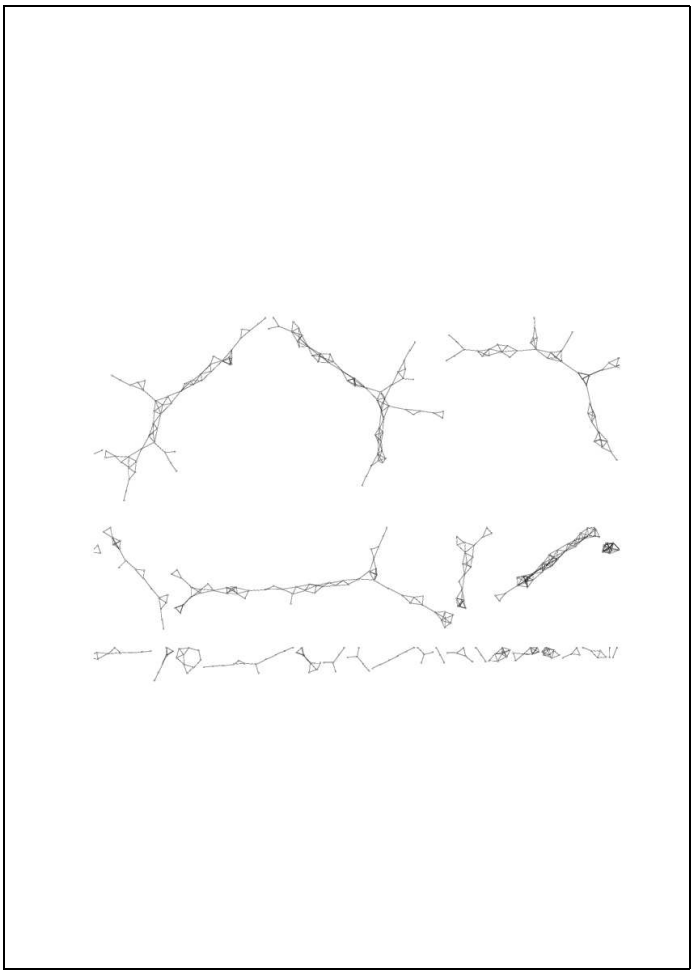


Figure 2.7: Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.60. The image is zoomed on several of the largest components.

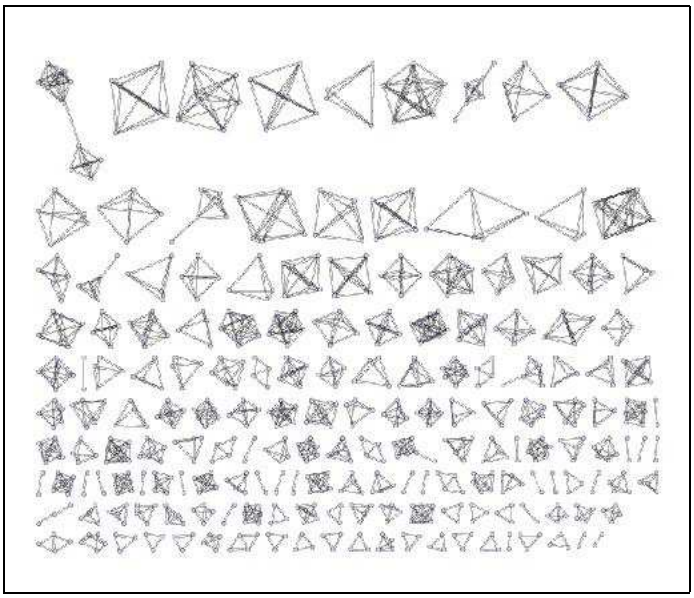


Figure 2.8: Random network on 1500 nodes, poisson degree distribution (z = 4), C = 0.97

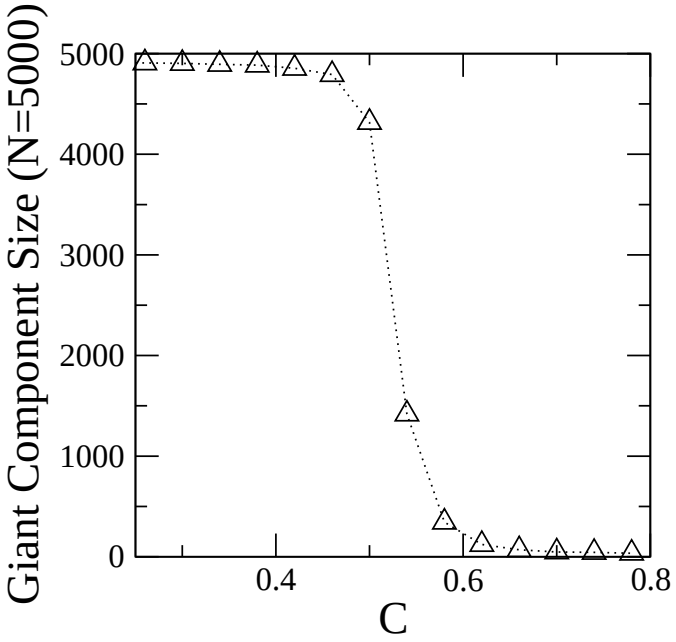


Figure 2.9: Size of the giant component versus the clustering coefficient in a poisson random network, $z = 3$. Each point represents the average of 40 trials.

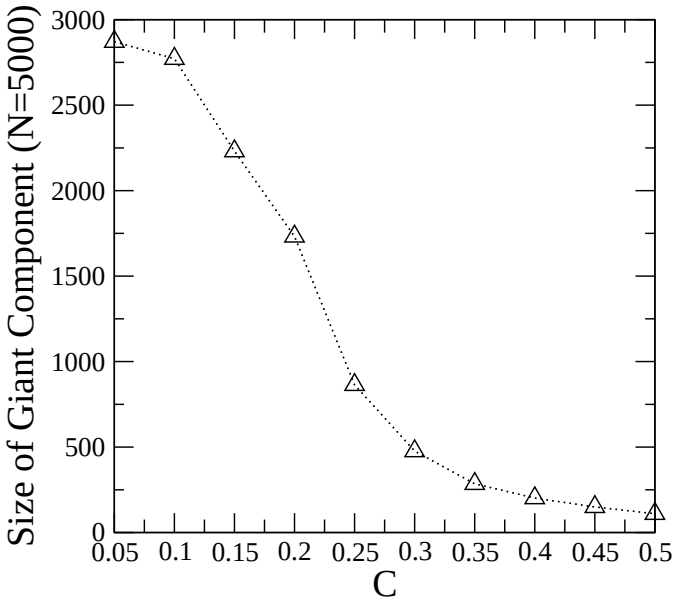


Figure 2.10: N=5,000 nodes. Power law with parameters $\kappa = 10$ and $\gamma = 2$. Each point represents the average of 40 trials. Compare this with 2.9. The phase transition is much less sharp than for the poisson random networks.

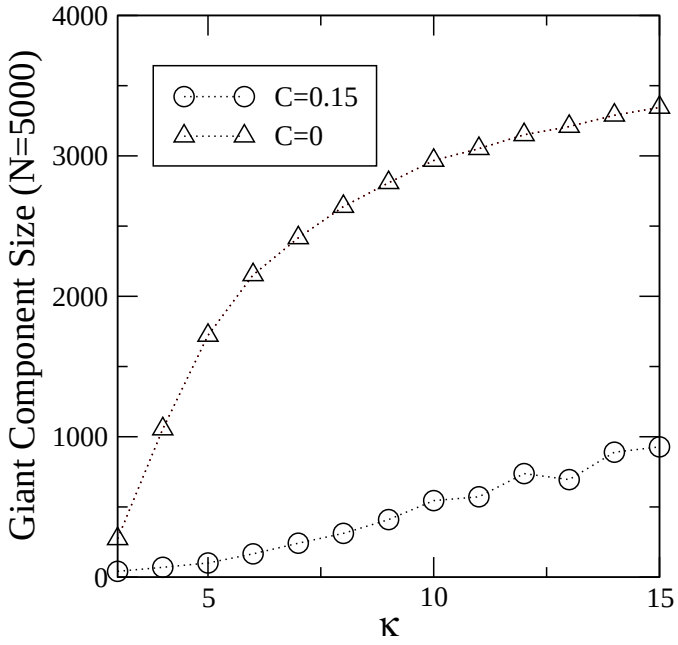


Figure 2.11: Two random networks are compared over a range of parameter values for the power law degree distribution with parameters κ and $\gamma = 2$. Each point represents the average of 40 trials.

as hubs for low-degree vertices, and consequently the network disconnects into many small components. Large components can be preserved under much higher clustering with distributions such as the poisson.

The phase transition also undergoes major changes with the introduction of clustering, although this effect seems to depend sensitively on the degree distribution. In figure 2.11 we see that the phase transition where a giant component forms is not significantly affected by the introduction of clustering for networks with power law degree distributions. In contrast to the poisson random networks, there is no sharp phase transition between the regime with a giant component and without. This bears some resemblance to percolation phenomena, where the phase transition disappears for true power-laws and an exponent of 2. But in figure 2.13 we see that the point of the phase transition was dramatically shifted forward for the poisson random network. It is somewhat surprising to observe the phase transition being shifted *forwards* as our algorithm features the introduction of degree assortativity into the network. Previous research has shown the tendency of degree assortativity to shift the point of the phase transition backwards [24].

2.3 Variations on the agorithm

We have proposed a very simple example of how network-growth, degree-assortativity and preferential attachment can be combined to generate networks with desirable properties. In fact, many features of this algorithm can be changed to give different and interesting results. It may be that some features of our algorithm are sub-optimal. Variations on this algorithm may be more effective at generating networks with the desired properties (e.g. a desired level of clustering, see section 2.7). There may be more effective ways to introduce degree assortativity, or

to form a list of nodes for preferential attachment. This paper is almost certainly not the final word on this subject.

While the present algorithm was being designed, numerous similar growth algorithms were tried. This section will outline some processes similar to what we have focussed on this paper.

2.3.1 Methods for generating degree assortativity

In our initial network growth experiments, we did not introduce any degree-assortativity at all. As mentioned above, degree assortativity plays an important part in our ability to form triads to a network.

The response of the size of the giant component to the input clustering parameter C_{input} was very different, and is shown in figure 2.12. The relationship is approximately linear, and should be contrasted with the sharp decline in the size of the giant component observed above at the phase transition C^* (fig. 2.9).

Another variation on degree assortativity concerns the formulation of p_{ij}^c. This is not the only "probability of compatibility" we can devise. An alternative is clear from the way our growth algorithm works.

Let *depth* refer to the distance of a node from the initial node in the current component of a growing network. Let $parents(i)$ denote the set of nodes at a lower *depth* than node i which are connected to node i. $|parents(i)|$ will be the number of parents node i possesses. Let $descendants(i)$ denote the set of nodes connected to node i which are also at a strictly greater depth than node i. In practice, a descendant of node i can never be connected to a parent of node i. This is because the parents of node i have already had their free connections "reserved" by the time a descendant of node i is designating its own connections. Hence it is not most

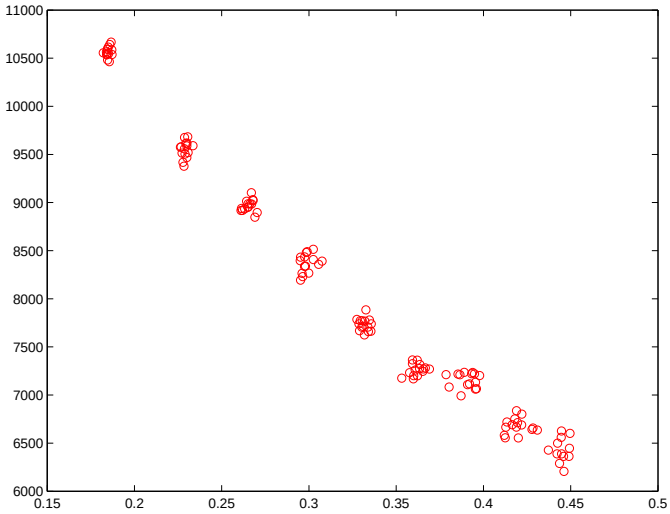


Figure 2.12: The size of the giant component is shown versus the input clustering parameter C_{input}. The network is Exponential(4), $n = 20000$

likely (sometimes even impossible) for a descendant of node i to connect to $C\delta(i)$ of i's neighbors. Rather the average number of triadic connections in common with i will be $C(\delta(i) - |parents(i)|$. The probability of compatibility between nodes i and j then becomes:

$$p_{ij}^{c} = \sum_{x=0}^{min\{d(i)-|parents(i)|,d(j)-|parents(j)|\}} p(\tau_{ij} = x)p(\tau_{ji} = x)$$

This modified degree-assortativity was not used in the experiments reported here. However, code for using this version of degree assortativity, as well as all of the other experiments can be found at

http://www.people.cornell.edu/pages/emv7/clustering.

2.4 Methods for generating lists of potential triads

There are various systems of preferential attachment which can be defined for growth networks. So long as every connected triple in the network becomes a

triad with probability C, the input clustering parameter will correspond to the output clustering. Therefore our preferential attachment rule should encourage the creation of triads as uniformly as possible for all connected triples. Unfortunately, a perfect way of accomplishing this has yet to be devised.

Sometimes the fate of two or more triples depends on the allocation of a single connection. This occurs whenever there are two or more paths of length two to a node which is represented in the list *PotentialTriads*. In these cases we have achieved the best results by allowing such a node to have multiple occurances in *PotentialTriads* and therefore to form a triad with probability greater than C_{input}. This method was in fact used for the experiments reported in this paper.

Another problem concerns nodes which are two steps away, but which nevertheless have no free connections; hence a triad could never be formed with that node. We have had some success with a method which compensates for this problem. Every time such a node is encountered, a random node is chosen from the *ProspectiveTriads* list, and is re-added to the list, such that it occurs with probability greater than C_{input}. This goes someway to compensating with new triads for triads which never had a chance to exist.

2.5 Phase transitions

It is a necessary condition for a giant component to exist that if we pick a node at random, the average number of neighbors two steps away, s_2, exceeds the number of neighbors one step away, s_1 [23]. This is intuitive, since if it were not the case, the number of neighbors n steps away would decrease to zero on average, and the component would be finite in the limit of large network size.We can use this to approximate the point of the phase transition as clustering is varied in our random

networks. Formally, we will solve for the point where

$$s_1 = s_2 \tag{2.1}$$

The necessary condition (2.1) will not quite be a sufficient condition in the presence of clustering as described below. Thus, our solution will only be a lower bound on the point of the phase transition, but in practice, this will serve as an excellent approximation.

For the poisson degree distribution, the average number of nodes one step away is equal to the parameter of the distribution z, so we have $s_1 = z$. As is well known [27], the number of edges emanating from a node if we pick an edge at random and follow it to one of its ends is also z for the poisson degree distribution. Thus, in the absence of clustering we would have simply $s_2 = s_1 z = z^2$, where s_2 is the average number of nodes two steps away from a randomly chosen node.

In the presence of clustering, things become more complicated. Lets pick a node uniformly at random in the network and call this node v_0. A neighbor of this node, v_1 will have on average z connections not in common with v_0. Furthermore, there will be on average Cz triadic connections between v_0 and v_1 as each of those connections has a probability C of being a triad. We can simply deduct the triadic connections from s_2, so that we have

$$s_2 > z^2 - Cz^2 = z^2(1 - C) \tag{2.2}$$

There is not equality in equation 2.2 because there is an additional force limiting the number of second neighbors: Once two neighbors of v_0, say v_1 and v_1' share a triadic connection, it becomes more likely that a node two steps away from v_0, say v_2, is a common neighbor of both v_1 and v_1'. In fact, such connections exist with probability C. Then, the number of connections we should deduct from every

neighbor at distance two due to common connections of nodes at distance one is equal to C times the average number of triadic connections at distance one, or in other words z^2C^2. Thus, we have

$$s_2 = z^2 - Cz^2 - C^2z^2 = z^2(1 - C - C^2)$$

We can use this to solve for the critical z_C^* where a giant component forms given a level of clustering C:

$$z = z^2(1 - C - C^2) \tag{2.3}$$

The non-zero root of this equation is given by

$$z_C^* = \frac{1}{1 - C - C^2} \tag{2.4}$$

Note that when C=0, we retrieve the well known result that a giant component forms when $z = 1$ in the absence of clustering. Unfortunately, we can only say that this is a lower bound for the phase transition due to that the nodes at distance two are not identical to v_0. The number of outgoing connections from such nodes (to nodes not already counted) is less than $z - C^2z$ on average.

In figure 2.13 we have plotted the size of the giant component versus the parameter z for several levels of clustering. The vertical lines correspond to the phase transitions z_C^* as given by (2.4). We find good agreement between theory and simulation.

There is a singularity in (2.4) where $1 - C - C^2 = 0$. At this point, $C^* = 0.618$, the giant component disappears regardless of the average degree z of the degree distribution. C^* represents the critical level of clustering that can coexist in a network with a giant component.

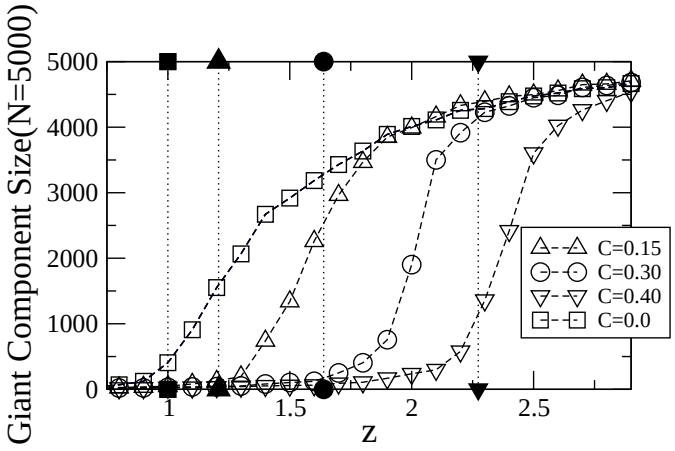


Figure 2.13: The size of the giant component is shown vs. z, the parameter of the poisson degree distribution, for four levels of clustering ($C = 0.0, C = 0.15, C = 0.30, C = 0.40$). The vertical lines indicate the point of the phase transition for each level of clustering predicted by equation 2.4

2.6 Finite size effects

During the execution of the algorithm, it occasionally happens that a node cannot find a suitable neighbor due to the absence of a node left in the network which has free stubs and the correct degree to satisfy the degree assortativity requirements. This imperfection is due to the finite size of the network. In the limit of large size, it would always be possible to find a scale such that every node can find just the right profile of neighbors with the right degree. There is no perfect way to deal with such discprepancies. For the simulations used in this article, we have simply truncated the degree of that node so that it does not have to seek a new neighbor. Even with networks of only 5000 nodes, the number of corrections made is quite small.

Figures 2.14 and 2.15 show the effects of network size and clustering on the

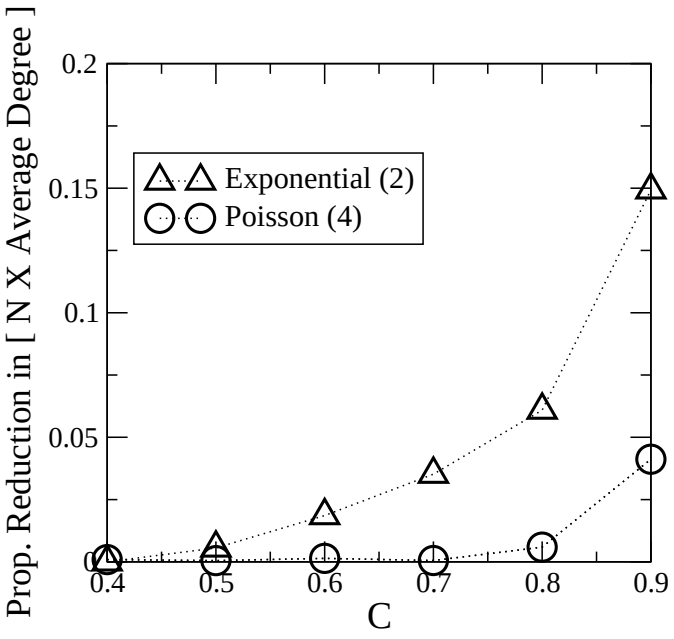


Figure 2.14: The percentage reduction in the number of "stubs" is shown versus the Clustering Coefficient for two networks: (i) Poisson degree distribution with parameter = 4, (ii) Exponential degree distribution with parameter = 2. N=5000 for both networks. Each point is based on the average of 20 trials.

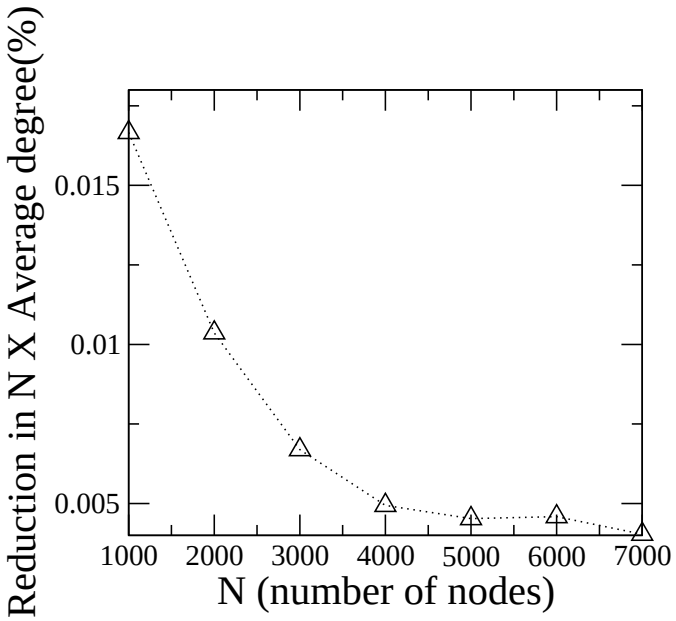


Figure 2.15: The percentage reduction in the number of stubs is shown versus the network size. The network has a Poisson degree distribution with parameter = 4, C = 0.80. Each point is based on the average of 20 trial networks.

amount of degree-corrections made by the algorithm. Figure 2.14 shows the effects of clustering on the number of corrections made for two networks. Note that the total number of stubs in the network is equal to the average degree of the nodes times the population size. The corrections made is shown as the proportional reduction in the number of stubs. Even at 90% clustering, the poisson random network only undergoes less than 5% reduction in its stubs.

Figure 2.15 shows the effects of network size on the number of corrections made. As expected, the number of corrections drops with the number of nodes in the network. For 7000 nodes and 80% clustering, a poisson random network undergoes less than a 0.4% reduction in its stubs.

2.7 Dependence of the clustering coefficient on input parameter C_{input}

We have demonstrated a random network model which can generate any desired level of clustering for any degree distribution. Getting a desired level of clustering C is not always as simple as setting the parameter $C_{input} = C$. In general the input clustering will be very close to the output clustering, though there are sometimes systematic differences. Figure 2.16 shows the value of the clustering coefficient achieved over a broad range of values of C_{input} for a Poisson random network. Although the C values do not always fall on the diagonal, they nevertheless cover the full spectrum of $C = 0$ to $C = 1.00$ making it possible to achieve any desired level of clustering.

It would be desirable for the input clustering to correspond exactly to the output clustering. The causes of the discrepancy are not fully understood as of the writing of this manuscript, but are probably related to innacurate degree-assortativity and improperly allocated prospective triad lists. Improving the algorithm so that C_{input} more closely corresponds to C would be worthy subject for future research.

2.8 Implications for sociology

The statistical properties of large social networks have been neglected by most social networks researchers in favor of the study of small networks which feature complete information about nodes and ties. This has begun to change in recent years as researchers from other disciplines have made great strides in the mathematics of large random networks– discoveries with direct applications to social net-

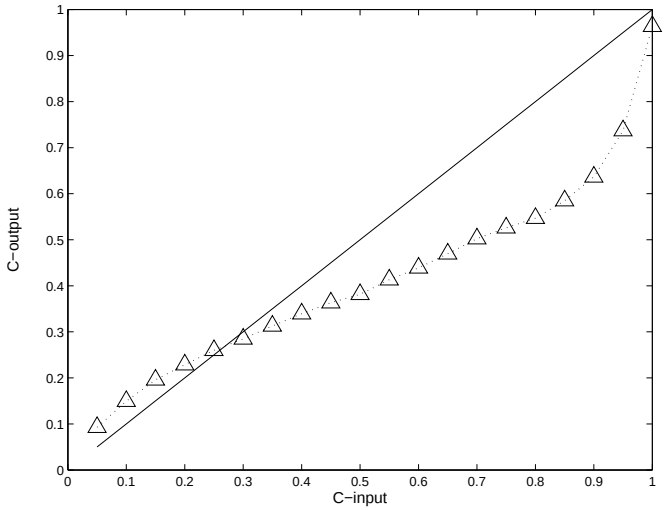


Figure 2.16: The clustering realized versus the input clustering parameter C_{input}. The random network has a poisson degree distribution with $z = 8$. $N = 2500$.

works. Indeed these advances were largely stimulated by a sociological question, the small-world problem, which was expertly investigated by Duncan Watts, an applied mathematician-turned sociologist. Now the methods developed by mathematicians and physicists are returning home to sociology where they may find new applications and facilitate our understanding of a broad range of large social networks, everything from markets and supply chains to internet-dating communities [15].

The present work aims to be a part of this quickly growing literature on large, complex social networks. From the very beginning of this literature– Duncan Watt's investigation of the small-world problem–transtivity of network connections has been a primary feature of interest. Duncan Watts explained how high transitivity can co-exist with short average path length. This was accomplished with a simple network model which featured random connections and transitivity which was built into a specified lattice topology and a constant degree distribution.

Watts did not, however, have a network model which allowed him to smoothly interpolate between various levels of clustering for any degree distribution. One significant aspect of this research is that it allows sociologists to explore broad ranges of clustering with realistic degree distributions. The degree distribution can even be taken directly from empirical data.

Another aim of this paper is to bring recognition to the multidude of mechanisms for injecting desired topologies into large random networks. Indeed, social networks researchers have been developing network models which feature transitivity for more than a decade [37]. In more recent years, *exponential random network* models have gained a strong foothold in the discpline. Network growth models have received less attention, and perhaps should receive more. Growth models are very flexible in the range of topologies they can produce. They are also suggestive of the mechanisms which produce the topologies we observe. For example, we have demonstrated that network growth and degree-assortativity coupled with preferential attachment to neighbors-of-neighbors is *alone* capable of generating large amounts of clustering.

Finally, a major contribution of this research to sociology is to clarify the relationship between transitivity and the connectivity of social networks. We have shown how increasing transitivity decreases the size of the giant component. Furthermore, there is an upper bound to transitivity, beyond which a giant component will not exist in a random network. It is unlikely that transtivity reaches such extremes in large social networks, as connectivity is an important feature to most of its constituents.

2.9 Discussion

We have presented a method for generating random networks which unite two frequently modeled topological features– clustering and the degree distribution.

Random network models can serve several important purposes. First, they can serve as a null hypothesis about the structure of a real-world network. Significant deviations in the structure of the real-world network from a corresponding random graph indicate that there are more forces at work shaping the network than are being accounted for in the random graph model. These deviations can then motivate further inquiry into the forces shaping real-world networks [27].

Secondly, real-world networks are very often of a scale that it is impossible to map them entirely. Various network sampling techniques have been devised to estimate features of the network topology in the absence of data on the entire network [14, 31, 6]. Given reliable estimates about network topology, a random network can then be generated which reproduces this topology. The random network may be used as a stand-in for modeling various dynamic models on networks.

Lastly, the family of random networks we have presented here enables the exploration of a huge parameter space for models on networks. There are a growing number of models which describe dynamic processes on networks. Examples are models of diffusion processes, such as models of epidemics [4, 20, 8], models of fads [35], the spread of rumors [38], the spread of innovations [12], and the migration of species among connected habitats [11]. Other models explore interactions among nodes embedded in a network. Examples include spin-glasses, kuramoto oscillators, and disordered neural networks [18]. There are many applications for exploring the effects of clustering and degree distributions on these and other models.

REFERENCES

[1] R. Albert and A.L. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47, 2002.

[2] R. Albert, H. Jeong, and A.L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:6794, 2000.

[3] L.A.N. Amaral, A. Scala, M. Barthélémy, and H.E. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA*, 97:11149–11152, 2000.

[4] L. Ancel-Meyers and M.E.J. Newman. Applying network theory to epidemics: control measures for mycoplasma pneumoniae outbreaks. *Emerging Inf. Dis.*, 9:204, 2003.

[5] L. Barabasi. *Linked*. Perseus, Cambridge, 2002.

[6] S.P. Blythe, C. Castillo-Chavez, and G. Casella. Empirical methods for the estimation of the mixing probabilities for socially structured populations from a single survey sample. *Math. Pop. Stud.*, 3:199–225, 1992.

[7] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Muñoz. Scale-free networks from varying node intrinsic fitness. *Phys. Rev. Lett.*, 89:258702, 2002.

[8] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Phys. Rev. E*, 86:3682, 2001.

[9] J. Davidsen, H. Ebel, and S. Bornholdt. Emergence of a small world from local interactions: Modeling acquaintance networks. *Phys. Rev. Lett.*, 88:128701, 2002.

[10] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, 85:4633, 2000.

[11] S. Ellner. Effects of successional dynamics on metapopulation persistance. *Ecology*, 84:882–889, 2003.

[12] X. Guardiola, A. Diaz-Guilera, C.J. Perez, A. Arenas, and M. Llas. Modeling diffusion of innovations in a social network. *Phys. Rev. E*, 66:026121, 2002.

[13] J.L. Guillaume and M. Latapy. A realistic model for complex networks. Preprint cond-mat/0307095, 2003.

[14] D. Heckathorn. Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49:11–34, 2002.

[15] P. Holme, C.R. Edling, and F. Liljeros. Structure and time evolution of an internet dating community. *Social Networks*, 26:155–174, 2004.

[16] P. Holme and B.J. Kim. Growing scale-free networks with tunable clustering. *Phys. Rev. E*, 65:026107, 2002.

[17] E. M. Jin, M. Girvan, and M. E. J. Newman. Structure of growing social networks. *Phys. Rev. E*, 64:046132, 1998.

[18] B.J. Kim. Performance of networks of artificial neurons: The role of clustering. *Phys. Rev. E*, 69:045101(R), 2004.

[19] P.L. Krapivsky and S. Redner. Organization of growing random networks. *Phys. Rev. E*, 63:066123, 2001.

[20] M. Kretzschmar and M. Morris. Measures of concurrency in networks and the spread of infectious diseases. *Math. Biosciences*, 133:165–195, 1996.

[21] J.F.F. Mendes and A.N.Samukhin. How to generate a random growing network. Preprint cond-mat/0206132, 2002.

[22] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. Uniform generation of random graphs with arbitrary degree sequences. Preprint cond-mat/0312028, 2003.

[23] Molloy and Reed. A critical point for random graphs with a given degree sequence. *Random Struct. and Algorithms*, 6:161, 1995.

[24] M.E.J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, 2003.

[25] M.E.J. Newman. Properties of highly clustered networks. *Phys. Rev. E*, 68:026121, 2003.

[26] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[27] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, 2001.

[28] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99:2566–2572, 2002.

[29] E. Ravasz and A.L. Barabási. Hierarchical organization in complex networks. Preprint cond-mat/0206130, 2002.

[30] E. Ravasz, A.L. Somera, D.A. Mongru, Z. Oltvai, and A.L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 2002:1551–1555, 2002.

[31] M. Salganik and D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193–240, 2004.

[32] S.N.Dorogovtsev and J.F.F. Mendes. *The evolution of networks: from biological nets to the Internet and WWW*. Oxford University Press, Oxford, 2003.

[33] D. Stauffer and A. Aharony. *Introduction to percolation theory*. Taylor & Francis, London, 1992.

[34] D.J. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, 1999.

[35] D.J. Watts. A simple model of global cascades on random networks. *Proc. Nat. Acad. Sci. USA*, 99:5766–5771, 2002.

[36] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[37] J. Weesie. A transitive random network model. *Social Networks*, 84:363–386, 1989.

[38] D. Zanette. Dynamics of rumor-propagation on small-world networks. *Phys. Rev. E*, 65:041908, 2002.

CHAPTER 3

TOMOGRAPHY OF RANDOM SOCIAL NETWORKS

Random network models have a long history in the social networks literature. Rapoport et. al. were the first to propose random graphs as models of social networks [22, 21, 20], while simultaneously the basic theory of random graphs was established in the mathematics literature by Erdős et. al [5]. Thereafter, periodic efforts were made to specify with greater detail the random or statistical nature of social networks, for example with the *biased random net* theory of Frank [8], Skvoretz [24], Fararo [6, 7], and others.

More recently, significant contributions have been made by statistical physicists, especially regarding the aggregate statistical attributes of networks [17, 19, 15]. The degree distribution has been shown to be one of the most important features of a network in determining network structure. Consequently, random networks with specified degree distributions have been proposed as a model of large, complex social networks [18, 10, 14, 16].

In this article, we describe techniques for revealing subtle aspects of network structure, taking as given a certain degree distribution. Our method relies on *network tomography* [12], the idea of mapping out a network layer by layer from a single node. The method is described in section 3.1 below.

The appropriateness of the random graph model must vary from population to population. Certainly a degree distribution does not determine the overall structure of a network. It is possible for a network with a given degree sequence to have extreme differences from a corresponding random network [16, 27, 25]. But even in such cases, differences are likely to be informative, suggesting unique mechanisms that move a network away from the random regime.

This work has implications for networks sampling, the study of diffusion and mathematical epidemiology, as well as other dynamic processes on networks. All of these problems involve the marriage of network structure with network dynamics. To answer dynamical questions, it is desirable to specify network structure with greater precision. Unfortunately, even in random networks of the type studied here, namely semi-random networks with given degree distributions, there are many topological questions which remain unanswered. We will focus on two: 1. How many individuals are there at any distance from a given node? 2. Among all nodes at a given distance, what is the degree distribution among those nodes? Example applications are further described in section 3.4.

3.1 Network tomography

In all that follows, we assume a network size n, and a degree distribution p_k (The probability of a node being degree k is p_k). Multiple connections and loops are allowed, however it should be noted that such connections are exceedingly rare for large n. Our networks are undirected. Connections within the network are entirely random but for these constraints.

Having constructed such a network, we can play the following thought experiment. Pick a node, v_0 uniformly at random within the giant component of the network[1]. We will call v_0 the *seed*. This node will have a degree ≥ 1, and a number of neighbors at distance one. Those nodes in turn will have a degree distribution specific to themselves, and a number of connections to other nodes at distance two from v_0. We can continue in this way, eventually breaking the entire giant

[1]A *component* in a network is a maximal set of nodes such that there exists a path between any two of them. A *giant component* is a component which occupies a fraction of the nodes in the network in the limit of large network size.

component into disjoint sets defined by the distance from our seed. Some nodes may not be enumerated in this way, in which event they fall outside of the giant component.

What we just described is the basic premise of *network tomography.*Network tomography, originally described in [12], is a method for revealing the structure of a random network by exploration, layer by layer, from a single starting node.

Now we can ask a host of questions with consequences for the structure of the network as a whole:

- How many nodes are there at distance l from the seed v_0?
- What is the degree distribution within each layer?
- What is the size of the giant component?
- What is the degree distribution within the giant component versus outside the giant component?
- What is the expected centrality of a seed v_0 picked at random in this way? What about the centrality of a degree k node?

All of these questions can be answered as outlined below. The method is shown schematically in figure 3.1.

Let S_l be the number of connections originating from layer l. For example, for $l = 0$, S_0 is the degree of v_0. Let R_l be the number of connections from layer $l-1$ to layer l. Finally, let T_l be the number of connections originating from nodes *outside* of layers $m \leq l$.

Let $S_0 = z_0$ where z_0 is the average degree in the giant component of the

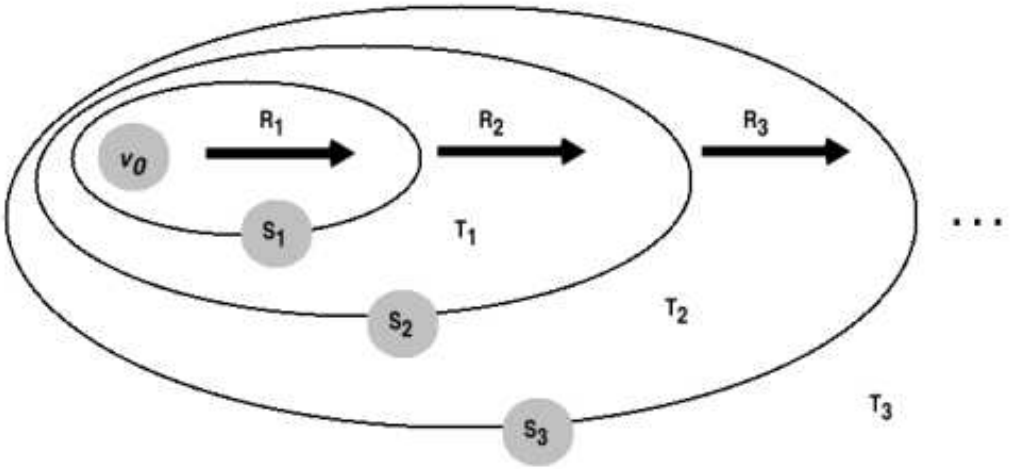


Figure 3.1: This diagram illustrates the tomographic method detailed in the text. Starting from a single node v_0 we recursively explore nodes at distance l from v_0. R_l is the number of connections going to layer l from layer $l-1$. S_l is the number of connections to nodes in layer l. T_l is the number of connections not connected to nodes in layer l or less. The importance of these quantities is explained in the text.

network[2]. $T_0 = nz - z_0$, where z is the average degree in the network as a whole, and $R_0 = 0$. To continue mapping out the network, we need a recurrence relation on these quantities:

$$\begin{aligned} S_{l+1} &= f_S(S_l, T_l, R_l) \\ T_{l+1} &= f_T(S_l, T_l, R_l) \\ R_{l+1} &= f_R(S_l, T_l, R_l) \end{aligned}$$

To proceed further, and determine the exact form of $f_\cdot(\cdot)$, we will need to draw on a technique widely employed in the complex networks literature, the *probability generating function*. Probability generating functions have found numerous applications to the study of complex networks. The first examples were given in [17, 18]. A good general reference to generating function methods is [30], and applications of generating functions to branching processes are given in [9] and [2].

Probability generating functions are created by transformation of discrete probability distributions into the space of polynomials. We will need just one generating function corresponding to our degree distribution:

$$g(x) = p_0 + p_1 x + p_2 x^2 + \cdots \tag{3.1}$$

Frequently we find that generating functions converge to simple algebraic functions, in which cases we can perform any operation on the algebraic version of the generating function instead of the series expansion. This constitutes one of the primary uses of probability generating functions.

In the examples that follow we will concern ourselves with two easy to study degree distributions:

[2]We can choose any degree for our seed, though some of the statistics we derive will be dependent on this parameter.

1. Poisson. This is the degree distribution of classical random graphs as studied by the Erdős and Rapoport among others. $p_k = \frac{z^k e^{-z}}{k}$. This is generated by

$$g(x) = e^{z(x-1)} \tag{3.2}$$

2. Exponential. $p_k = (1 - e^{-1/z})e^{-k/z}$. This is generated by

$$g(x) = \frac{1 - e^{-1/z}}{1 - xe^{-1/z}} \tag{3.3}$$

See [15] for a derivation of these generating functions.

Returning to the tomography problem, consider the probability that a connection emerging from layer l will go to a node in layer $l+1$, given that the connection does not go to layer $l-1$. Since our networks are completely random, such a connection has uniform probability of going to any of the "stubs" originating from nodes in layers $m > l$, as well as stubs originating from nodes in layer l, minus those stubs which are already allotted to layer $l-1$. This gives us the following:

$$P_{l \to l+1} = \frac{T_l}{T_l + S_l - R_l}$$

For convenience, we now define the following quantity:

$$\alpha_l = \alpha_{l-1} \frac{T_l}{T_l + S_l - R_l}$$

This is the probability of a conjunction of events, namely that a connection goes to a node outside of layer l, given that the connection has not attached to layers $m < l$.

Note that the probability that a degree k node lies outside the first l layers is the probability that all k of the nodes connections go to other nodes outside of layers $m \leq l$. This is simply α_{l-1}^k.

Now it can be asked: What is the average degree of a node outside of layers $m \leq l$? We have

$$< k >_{T_l} = \sum_k \alpha^k \; p_k \; k/c \tag{3.4}$$

where c is the appropriate normalizing constant:

$$c = \sum_k \alpha^k p_k$$

The value of our generating function approach is now apparent, as we can easily express the above in terms of our generating function $g(x)$:

$$< k >_{T_l} = n[\frac{\mathrm{d}\ \mathrm{g}(\alpha_\mathrm{l}\mathrm{x})}{\mathrm{dx}}]_{x=1}/g(\alpha_l) = \alpha_l g'(\alpha_l)/g(\alpha) \tag{3.5}$$

By similar reasoning, the total number of connections originating from nodes outside of layer $l+1$ is:

$$T_{l+1} = n[\frac{\mathrm{d}\ \mathrm{g}(\alpha_\mathrm{l}\mathrm{x})}{\mathrm{dx}}]_{x=1} = n \; \alpha_l \; g'(\alpha_l) \tag{3.6}$$

Once this is known, S and R follow easily. S is equivalent to the change in the number of connections between two adjacent layers. R will be the expected number of connections going between two adjacent layers. We have:

$$\begin{aligned} S_{l+1} &= T_l - T_{l+1} \\ R_{l+1} &= S_l \frac{T_l}{T_l + S_l - R_l} = S_l \alpha_l / \alpha_{l-1} \end{aligned}$$

This recurrence relation can be solved to any desired depth. Below it will be shown that many interesting quantities can be computed from the sequences of S,T, and R. [3]

[3]It is worth noting that the recurrence relation on S,T, and R can be simplified to a recurrence relation on just two variables, due to that S is not a function of

3.1.1 Descriptive statistics

Let's return the questions from section 3.1. With the simple recurrence relation 3.6 and 3.7 we can now characterize many feature of our network. Once a sequence of values of S_l, R_l, and T_l have been computed, it is quite simple to determine many things about the structure of our network by plugging in the appropriate values into our generating functions.

Of foremost importance is the size of each layer, that is the number of nodes at some distance from our seed. We know that the probability of a degree k node being outside layer l is α_l^k. Then the probability of a degree k node being within layer l is $\alpha_{l-1}^k - \alpha_l^k$. So, choosing a node at random, the probability of that node being in layer l will be $\sum_k p_k(\alpha_{l-1}^k - \alpha_l^k)$. Translating this into our generating function language, and multiplying by the population size n, we have

$$n_l = n(g(\alpha_{l-1}) - g(\alpha_l)) \tag{3.7}$$

The size of the giant component is even easier to derive. Let $\alpha_\infty = \lim_{l\to\infty} \alpha_l$ [4]. This is the probability that a connection goes to a node at distance infinity from the seed, or in other words is outside of the giant component. The probability itself. Specifically, by eliminating S, we get

$$T_{l+1} = n\frac{\frac{d}{dx}[g(\alpha_{l+1}x)]_{x=1}}{g(\alpha_l)}$$

$$R_{l+1} = \frac{T_l(T_{l-1} - T_l)}{T_{l-1} - R_l}$$

and

$$\alpha_l = \alpha_{l-1}\frac{T_{l-1}}{T_{l-2} - R_{l-1}}$$

[4]It is interesting to note that α_∞ corresponds to the probability of a connection not being to the giant component, u, as derived by Newman et al. in [17]. The way that this quantity is computed is somewhat different.

that a degree k node is outside the giant component is then α_∞^k. Following similar reasoning as above we find the size of the giant component to be

$$n_{gc} = n(1 - g(\alpha_\infty)) \tag{3.8}$$

As we move outward from our seed, we find that the degree distribution changes within each layer of the network. Initially the average degree tends to increase, as nodes are connected to with probability proportional to degree. But quickly high degree nodes are exhausted, and the average degree within a layer decreases sharply.

In the l'th layer the probability of a node being degree k given by

$$p_{k;l} = \frac{p_k}{c}(\alpha_{l-1}^k - \alpha_l^k) \tag{3.9}$$

$$= \frac{p_k}{c}\Big(1 - \frac{T_l}{T_l + S_l - R_l}\Big)\alpha_{l-1}^k \tag{3.10}$$

$$= \frac{p_k}{c}\frac{S_l - R_l}{T_l + S_l - R_l}\ \alpha_{l-1}^k \tag{3.11}$$

where c is the appropriate normalizing constant for the degree distribution. When α is close to zero, it dominates the above expression, and thus the distribution converges to a power law as we move away from the seed. Of course, if p_k decays faster than a power law (e.g. exponentially) then the distribution will theoretically not have the "fat tails" characteristic of power-laws for large k. This happens regardless of the degree distribution of the network as a whole.

Using identical reasoning as we used to determine the number of nodes in layer l, we can determine the generating function for the degree distribution in layer l.

$$g_l(x) = \frac{g(\alpha_{l-1}x) - g(\alpha_l x)}{g(\alpha_{l-1}) - g(\alpha_l)} \tag{3.12}$$

Note that $g(\alpha_{l-1}) - g(\alpha_l)$ is in the denominator to normalize the distribution.

The degree distribution outside of the giant component is similarly easy to derive:

$$g_{gc^c}(x) = \frac{g(\alpha_\infty x)}{g(\alpha_\infty)} \tag{3.13}$$

And the degree distribution within the giant component is the complement:

$$g_{gc}(x) = \frac{g(x) - g(\alpha_\infty x)}{1 - g(\alpha_\infty)} \tag{3.14}$$

An important sociological consideration is the mean path length and the associated closeness centrality statistic [29, 11]. Having chosen a seed, we can compute the average distance to other nodes in the network using the quantities calculated above:

$$m_c = \sum_{l \geq 1} \frac{l \times n_l}{n_{gc}} \tag{3.15}$$

This can be considered the expected closeness centrality of a degree z_0 node in the network, where z_0 is the degree of our seed.

3.2 Theoretical Examples

The reader may find it helpful if we illustrate the preceding ideas with a few simple, idealized examples.

Many social networks fall into one of two regimes. The simplest case is for the degree distribution to be relatively homogeneous, as occurs when individuals connect to one another with uniform probability. This leads to the classical random networks such as those studied by Rapaport and Erdős. These are characterized by a symmetric, unimodal distribution, namely the Poisson generated by equation 3.2. In the second regime, we find that a minority of individuals act as "hubs" for the network, thereby accounting for the great majority of connections in the network [3]. This leads to highly skewed degree distributions such as power-laws

and simple exponentials. Although highly idealized, both of these simple cases may have something to teach us about the structure of real social networks.

We have explored both Poisson and Exponential networks using simulation and the tomographic methods discussed above. Consider the Poisson degree distribution, with generating function 3.2. Let $n = 50000$.

By combining equations 3.2 and 3.11 we find that the degree distribution in layer l is generated by

$$g_l(x) = \frac{e^{z(\alpha_{l-1}x-1)} - e^{z(\alpha_l x-1)}}{e^{z(\alpha_{l-1}-1)} - e^{z(\alpha_l-1)}} \tag{3.16}$$

$$= e^{z\alpha_{l-1}(x-1)} \frac{1 - e^{z\alpha_{l-1}x(\gamma_l-1)}}{1 - e^{z\alpha_{l-1}(\gamma_l-1)}} \tag{3.17}$$

where $\gamma_l = T_l/(T_l + S_l - R_l)$.

It can be verified that this satisfies the requirements for a probability generating function, namely that it has a series expansion, and that $g_l(1) = 1$. Figure 3.2 shows the degree distribution for $z = 3$ at various layers. The solid lines represent the theoretical solutions given by 3.11, and the points, where present, mark the results of simulation. 40 networks of size $n = 50000$ and with Poisson degree distribution, $z = 3$ were generated. For each network 20 seeds were chosen independently, and the network was mapped out from each. Averaging these simulations yield the data points shown.

Furthermore we can explore how the network changes its structure as the mean of the degree distribution, z, is swept over a range of values. Figure 3.3 shows the results of one simulation where $z = 1.25, 3, 5$ and $n = 50000$ as before. The average number of nodes at various distances from a randomly chosen seed is shown. Dotted lines represent the results of simulations, while the solid lines represent the theoretical prediction. The dotted line above the theoretical prediction shows

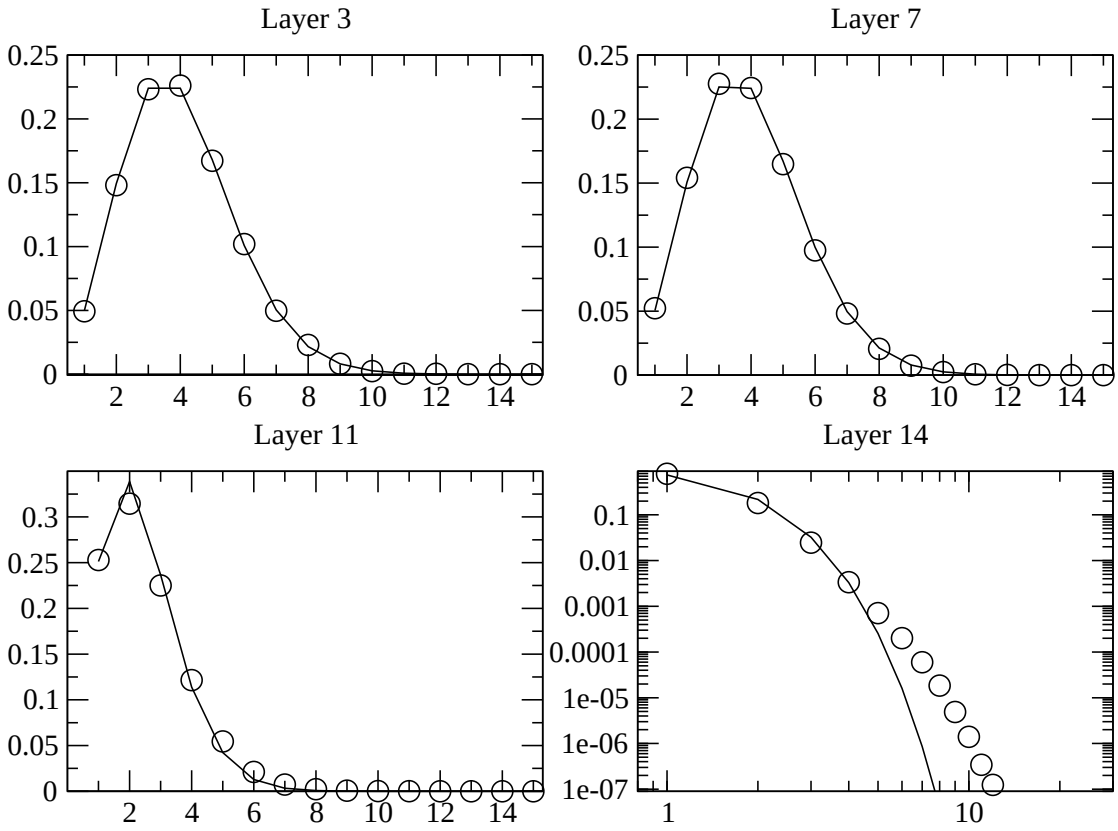


Figure 3.2: $n = 50000$, Poisson degree distribution, $z = 3$. Data points are the average of 40 generated networks with 20 trials per network. Solid lines represent the theoretical prediction given by 3.17.

the 90'th percentile among simulations. Likewise the dotted line below shows the 10'th percentile. It can be seen that our theory correctly captures the trend as we increase z from 1.25 to 5.

The theoretical prediction for figure 3.3 is derived by solving our generating function 3.2 and using 3.7. We find:

$$n_l = ne^{z(\alpha_{l-1}-1)}(1 - e^{z\alpha_{l-1}(\gamma_l-1)}) \tag{3.18}$$

where $\gamma_l = T_l/(T_l + S_l - R_l)$.

Figures 3.4 and 3.5 show identical experiments for the exponential degree distribution 3.3. The mathematics is somewhat more tedious for this case, so we omit it here.

Now viewing the results for the exponential and Poisson experiments, several things bear mention. As we observed above, the degree distribution converges to a skewed exponential or power-law as we move to higher layers in the network. This occurs despite the homogeneous degree distribution of the Poisson networks. In fact, our theory predicts an exponential tail for both of these distributions for high layers. However, we observe the "fat-tails" of power laws instead. This is most likely a finite-size effect.

The existence of hubs in the exponential networks lead to several interesting differences with the Poisson networks. It can be seen from the n_l experiments that the exponential has a narrower peak than the Poisson. As soon as a path is found from v_0 to a hub, the rest of the network can be reached in very few steps. It is also interesting that the degree distribution for the exponential random networks has its mode shifted rightward of 0 in the first several layers, thus making its distribution more reminiscent of the Poisson. This is yet another consequence of the existence of hubs in these networks; the higher mode bulge in these distribution represents

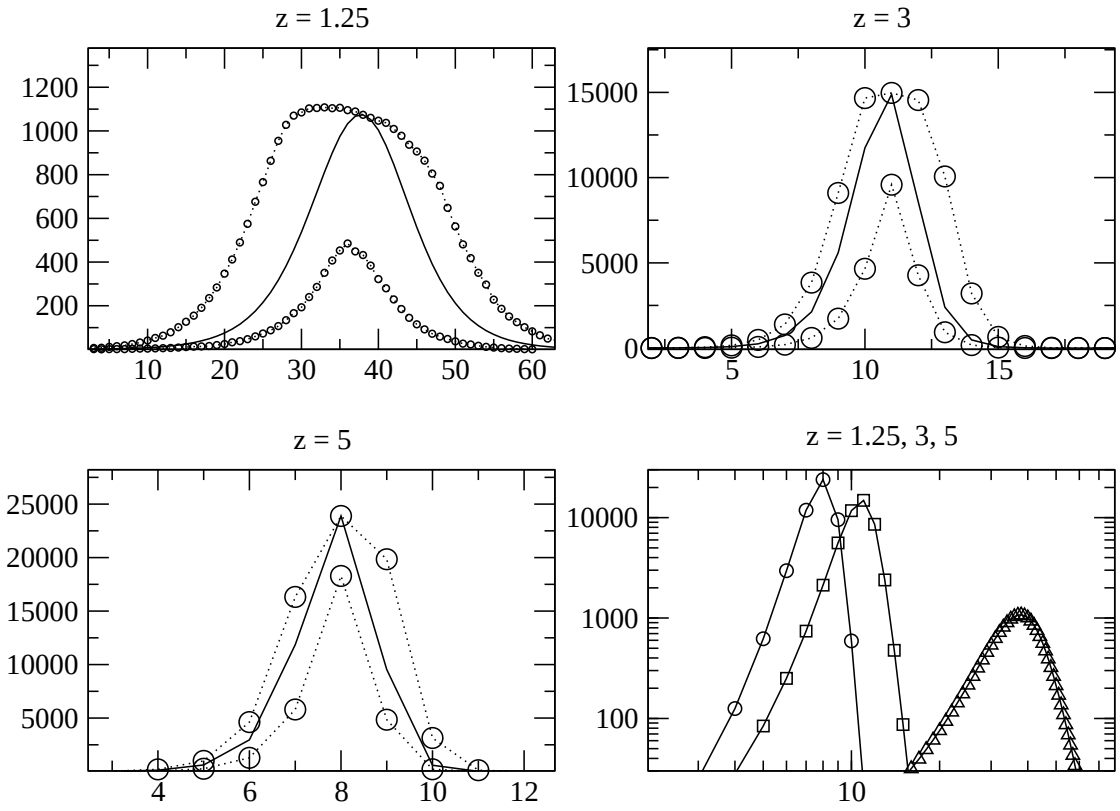


Figure 3.3: $n = 50000$, Poisson degree distribution, $z = 1.25, 3, 5$. Data points show the 10'th and 90'th percentile for 40 randomly generated networks with 20 trials per network. Solid lines represent the theoretical prediction given by 3.18.

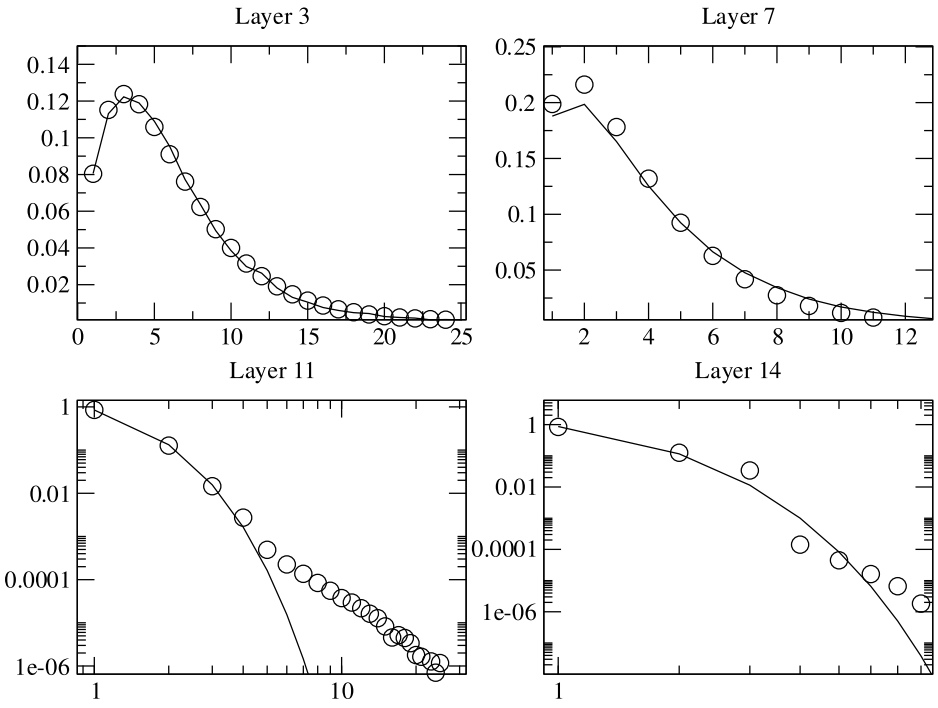


Figure 3.4: $n = 50000$, Exponential degree distribution, $z = 3$. Data points are the average of 40 generated networks with 20 trials per network. Solid lines represent the theoretical prediction given by 3.11.

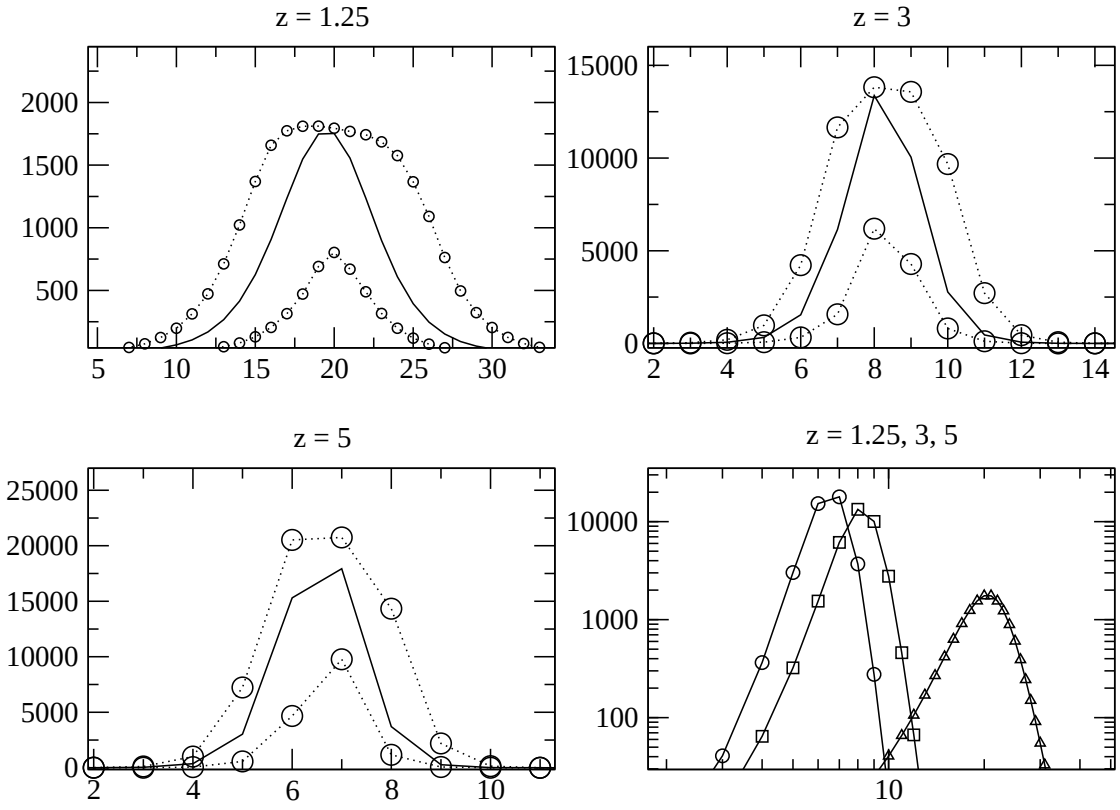


Figure 3.5: $n = 50000$, Exponential degree distribution, $z = 1.25, 3, 5$. Data points show the 10'th and 90'th percentile for 40 randomly generated networks with 20 trials per network. Solid lines represent the theoretical prediction given by 3.7.

the existence of higher degree hubs a short distance from v_0.

3.3 Email Network

The ideas presented here can be illustrated with a real social network. The network shown in figure 3.6 is the giant component for a one-day sample of email traffic for individuals at Cornell University. This includes a diverse collection of faculty, researchers, students and administrators. The communication linking them is correspondingly diverse, motivated by work, research and social affiliation.

Figure 3.6: The giant component from the Cornell email network. Connections in the network represent reciprocal communication within a 24 hour sampling frame. The nodes are color-coded. Blue nodes are faculty, red nodes are graduate students, green nodes are undergraduates, and yellow nodes are everyone else, mainly administrators. The network 2607 nodes and 4838 connections. The giant component consists of 1227 nodes.

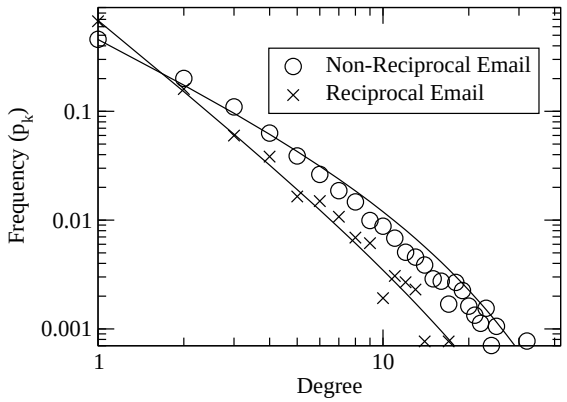


Figure 3.7: Degree distributions for the reciprocal and non-reciprocal email networks. Solid lines show a fit designed to match the average degree of the empirical distribution. The theoretical density is given by equation (3.19).

In communication networks such as these, it is very important to develop a sense of tie-strength between individuals, particularly for email networks, as a great deal of communication does not indicate a meaningful relationship, but merely the spread of cheap information (i.e. "spam"). Fortunately, there is an easy way to distinguish genuine social affiliation from simple information transfer. If persons in the network exchange emails in both directions within the 24 hour sampling frame, that is a strong indication that the conversants are well-acquainted and socially connected. We can then induce a subnetwork by including only those ties which are reciprocal.

In what follows, two networks will be considered. The first is the raw communication network, with no distinction made between reciprocal and non-reciprocal communication. For convenience, this will be referred to as the *R/NR* network. This network consists of 14216 nodes with 25040 connections. The giant component of the network occupies 13577 of the nodes (95.5%).

The second network consists only of reciprocal email connections and the nodes which have such connections. This will be called the R network. This network is much smaller, consisting of only 2607 nodes with 4838 connections. The giant component occupies 1227 nodes (47.1%).

The degree distributions for both the R and R/NR networks are shown in figure 3.7. Both distributions are evidently power laws, as they lie approximately on a straight line with log/log axes. The solid lines show a fit to these data of a power law density with exponential cutoff:

$$p_k = \frac{k^{-\gamma} e^{-k/\kappa}}{Li_\gamma(e^{-1/\kappa})}, k \geq 1 \tag{3.19}$$

where $Li_n(x)$ is the nth polylogarithm of x. To apply the tomographic theory, we need the generating function for this density. This is given by

$$g(x) = Li_\gamma(xe^{-1/\kappa})/Li_\gamma(e^{-1/\kappa}). \tag{3.20}$$

When applying the tomographic theory, it is possible to use the empirical degree distribution, but as the theoretical distributions appear to fit the empirical power laws very well, we will use the theoretical distributions instead. Figure 3.8 shows the stratum sizes predicted for the R/NR network using equation (3.7) (solid line). The dotted lines above and below the theoretical prediction are the actual 90'th and 10'th percentile stratum sizes from the R/NR network. The theory matches observations fairly well for the R/NR network. A very different situation is illustrated by figure 3.9, which shows the theoretical stratum sizes (solid line) alongside the mean stratum size for the R network (dotted line). There is clearly a great deviation between theory and observation. Nevertheless, this difference is instructive. The R network shows only strong ties, in contrast to the R/NR network which contains both strong and weak ties. Consequently, there are many more

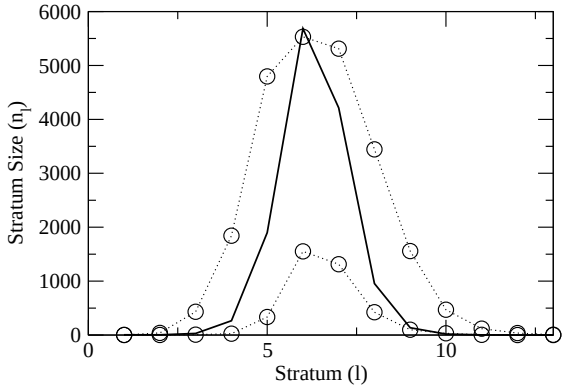


Figure 3.8: Theoretical (solid line) and empirical (dotted line) stratum sizes for the R/NR email network. This network includes both reciprocal and non-reciprocal communication within the 24 hour sampling frame. The upper dotted line represents 90'th percentile stratum sizes picking a *seed* from the network uniformly at random. The lower dotted line represents the 10'th percentile.

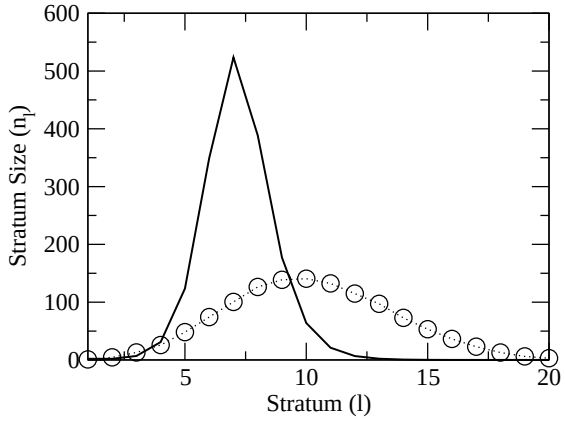


Figure 3.9: Theoretical (solid line) and empirical (dotted line) stratum sizes for the R email network. This network includes only reciprocal communication within the 24 hour sampling frame. The dotted line represents the mean empirical stratum size, selecting a *seed* from the network uniformly at random.

social micro-structures in the R network than would be expected in a pure random network. The *clustering coefficient*[5] , a measure of network transitivity, is much greater for the R network ($C = 7.4\%$) than for the R/NR network ($C = 1.86\%$). Of course, in a pure random network of these sizes, $C \approx 0$. Micro-structures such as these contribute to the deviations seen in figure 3.9 because they push the social network away from the pure random regime on which the network tomographic theory is based. As shown in [27], clustering has the effect of increasing mean path length and decreasing the giant component size. This is why a more elongated series of stratum sizes is observed in figure 3.9.

3.4 Discussion

The methods discussed here have relevance for disparate areas of networks research.

Consider the problem of *network sampling*– the utilization of social networks for surveying a population. Lately methods of chain-referral sampling have been proposed [23, 28] which model chain-referral samples as random walks on social networks. In general, little is known about the attributes of individuals reached after n steps of such a random walk. Tomographic methods may open a new window on the problem. We can now compute the expected properties of a node at a given distance from our starting point, as well as the probability that a random walk will be at that distance after a given number of steps. This allows us to answer questions such as

- How many different nodes could possibly be reached after n steps?

[5]The *clustering coefficient*, C, is defined as the ratio of the number of triads to the number of potential triads in a network: $C = 3\frac{N_\Delta}{N_3}$ where N_Δ is the number of triads in the network and N_3 is the number of connected triples of nodes. Note that in every triad there are three connected triples.

- What is the probability of the n'th node in a chain referral sample having degree k?
- What is the probability of being at distance l from our starting point after n steps?

It is beyond the scope of this paper to provide answers to these questions, but it is certainly possible using network tomography.

Another potential application is to the study of *network diffusion*– the study of dynamical processes which spread through a population via network connections. Examples include the adoption of innovations [26, 4] as well as the spread of information or rumors [1, 31]. The $\{n_l\}$ curves shown above are highly reminiscent of birth and death processes such as the spread of an epidemic through a population of susceptible individuals. In fact, the way we have mapped out our network from a single node is somewhat like the way an infectious agent may spread through a population from an initial infected. Previous research [13] has investigated the structural properties of diffusion of this sort, e.g. the proportion of the network that is ultimately occupied by infecteds. But it has been difficult to place a timescale on diffusion without resorting to computer simulation. It is hoped that progress will soon be made with the application of network tomography to these and related problems.

All of these results must be taken with the caveat that real networks may not be organized as simple random networks. As mentioned above, there is no guarantee that a real social network will exhibit the same sequences of n_l or $p_{k;l}$ as in the random regime. Extra forces can shape the network topology and push these statistics away from the pure random regime. These statistics can be thought of as something that help characterize the structure of the network, like a fingerprint

of its structure. When the statistics deviate from the random regime, it is an indication that unique and potentially interesting forces are affecting the network.

A simple example is furnished by the potential existence of greater than random *transitivity*(i.e. triadic closure), which can certainly affect the number of nodes at a given distance from our seed as well as the degree distribution at that distance [27]. However, with more study it may even be possible to adapt the tomographic method to account for transitivity and other non-random structures within social networks.

REFERENCES

[1] X. Guardiola aand A. Diaz-Guilera, C.J. Perez, A. Arenas, and M. Llas. Modelling diffusion of innovations in a social network. *Phys. Rev. E*, 66:026121, 2002.

[2] K. B. Athreya and P. Ney. *Branching Processes*. Springer, New York, 1972.

[3] L. Barabasi. *Linked*. Perseus, Cambridge, 2002.

[4] E.M.Rogers. *Diffusion of Innovations*. FF Shoemaker, New York, 1983.

[5] P. Erdős and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[6] T.J. Fararo. Biased networks and social structure theorems: part i. *Social Networks*, 3:137–159, 1981.

[7] T.J. Fararo. Biased networks and strength of weak ties. *Social Networks*, 5:1–11, 1983.

[8] O. Frank and D. Strauss. Markov graphs. *Journal of the American statistical association*, 81:832–842, 1986.

[9] T.E. Harris. *The Theory of Branching Processes*. Springer, Berlin, 1963.

[10] P. Holme, C.R. Edling, and F. Liljeros. Structure and time evolution of an internet dating community. *Social Networks*, 26:155–174, 2004.

[11] J.Scott. *Social Network Analysis: A Handbook*. Sage, London, 2nd edition, 2000.

[12] T. Kalisky, R. Cohen, D. ben Avraham, and S Havlin. *Complex Networks*, chapter Tomography and stability of complex networks. Springer-Verlag, New York, 2004.

[13] L.A. Meyers, B. Pourbohloul, M.E.J. Newman, D.M. Skowronski, and R.C. Brun-ham. Network theory and sars: Predicting outbreak diversity. *J. Theor. Biol.*, 232:71–81, 2005.

[14] M.E.J. Newman. Ego-centered networks and the ripple effect. *Social Networks*, 25:83–95, 2003.

[15] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[16] M.E.J. Newman and P. Juyong. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, 2003.

[17] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, 2001.

[18] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99:2566–2572, 2002.

[19] R. Pastor-Satorras, M. Rubi, and A. Diaz-Guilera, editors. *Statistical mechanics of complex networks*. Springer, Berlin, 2003.

[20] A. Rapoport. A contribution to the theory of random and biased nets. *Bulletin of Mathematical Biophysics*, 19:257–271, 1957.

[21] A. Rapoport. *Handbook of Mathematical Psychology*, chapter Mathematical models of social interaction. Wiley, New York, 1963.

[22] A. Rapoport and R. Solomonoff. Connectivity of random nets. *Bulletin of Mathematical Biophysics*, 13:107–117, 1951.

[23] M. Salganik and D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193–240, 2004.

[24] J. Skvoretz. Biased net theory: Approximations, simulations and observations. *Social Networks*, 12:217–238, 1990.

[25] T.A.B. Snijders. Accounting for degree distributions in empirical analysis of network dynamics. In R. Breiger, K. Carley, and P. Pattison, editors, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pages 146–161. The National Academies Press, Washington D.C., 2003.

[26] T.W. Valente. Social network thresholds in the diffusion of innovations. *Social Networks*, 18:69–89, 1996.

[27] E. Volz. Random networks with tunable degree distribution and clustering. *Phys. Rev. E*, 70:056115, 2004.

[28] E. Volz and D. Heckathorn. Probability based estimation theory for respondent driven sampling. Under Review, 2004.

[29] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, Cambridge, 1994.

[30] H.S. Wilf. *Generatingfunctionology.* Academic Press, Boston, 2nd edition, 1994.

[31] D. Zanette. Dynamics of rumor-propagation on small-world networks. *Phys. Rev. E*, 65:041908, 2002.

CHAPTER 4

SIR DYNAMICS IN POPULATIONS WITH HETEROGENEOUS CONNECTIVITY

Contact patterns constitute an important aspect of heterogeneity within a population of susceptible and infectious individuals. It has also been one of the hardest factors to incorporate into epidemiological models. Compartment models have been able to capture many aspects of population heterogeneity, such as with respect to heterogeneous susceptibility and infectiousness [27, 2, 9]. But compartment models can be inadequate with respect to population structure, especially when contact rates follow a steep and continuous gradient.

Network theory describes a population of susceptible and infectious individuals as nodes in a network [17, 26, 21, 13]. This has spawned a new category of epidemiological models in which epidemics spread from node to node by traversing network connections [24, 18, 20, 28, 8, 25]. Pure random networks with specified degree distributions have been proposed as realistic models of population structure. This case has the advantage of being well understood mathematically. The limiting behavior of epidemics spreading through random networks with a given degree distribution has been solved exactly [18, 20].

The network approach has the advantage that the mathematics of stochastic branching processes [29, 15, 4] can be brought to bear on the problem. This allows for precise descriptions of the distribution of outbreak sizes early in the course of the epidemic, as well as a solution for the final size of epidemics [18, 20].

A shortcoming of the network model has been that stochastic branching processes are inadequate to describe the explicit dynamical behavior epidemics. Thus the distribution of outbreak sizes are easy to solve for, yet the incidence curve,

Table 4.1: A summary of the nonlinear differential equations used to the describe the spread of a simple SIR type epidemic through a random network. The degree distribution of the network is generated by $g(x)$.

$\dot{\beta} = \alpha\ \mu\ p_W$
$\dot{\alpha} = -\alpha(r+\mu)p_W$
$\dot{W} = p_W(r\ n\ \alpha^2 g''(\alpha+\beta) - (r+\mu)(2W + n\ \alpha\ g'(\alpha+\beta)))$

that is the number of infecteds at a time t, has been difficult to derive. Simulation has been used in this case [11].

Heterogeneous networks make it difficult to derive differential equations to describe the course of an epidemic. Nevertheless, several researchers [5, 22, 23, 7, 10] have been successful modeling many of the dynamical aspects of network epidemics, particularly in the early stage where asymptotically correct formuli for disease incidence are now known. We improve upon these results by presenting a system of nonlinear differential equations which can be used to solve for the complete incidence curve, as well as other quantities of interest. We treat the simplest possible case of the SIR dynamics with constant rate of infection and recovery. Section 4.1 describes the model. Several examples are given in section 4.3.

4.1 Intuitive model specification

We investigate undirected random networks with specified degree distributions[1][26]. Let p_k be the probability of a node having a degree k. As in previous research we

[1]The *degree* of a node in a network is the number of connections to that node. The *degree distribution* is a discrete probability density over the positive integers describing the probability of realizing a given degree.

will make great use of the probability generating function (PGF) corresponding to the degree distribution.

Although widely employed in the probability theory and the study of stochastic branching processes, generating functions are less familiar to those working in mathematical epidemiology (but see[6, 12, 1, 3]). The utility of PGF's for the current investigation cannot be understated. Consider the degree distribution *among susceptibles* at a given time t. As an epidemic progresses, more highly connected nodes, often called "hubs", will be preferentially culled from the population of susceptibles. Thus the degree distribution among susceptibles will evolve as the epidemic progresses. Our approach will be to keep track of the evolution of this distribution by careful application of parameters to the PGF. This will ultimately allow us to find the number of infecteds at any given time.

Given a degree distribution, we define the probability generating function $g(x)$ as

$$g(x) = p_0 + p_1 x + p_2 x^2 + p_3 x^3 + \cdots \tag{4.1}$$

In most cases this series will converge to an algebraic function, in which case any operation to be done on the PGF can be done on the simple algebraic form. The series form can be retrieved by Taylor expansion. The degree distribution is a parameter of the model, so g must be well-defined. Several examples for common distributions are given in section 4.3. The results given below hold for any degree distribution.

It will be helpful to the reader if several examples are provided to illustrate the utility of PGF's. Generating functions allow us to manipulate probability densities using simple algebraic operations. For example, if we were to draw two realizations of a random variable X with generating function $g(x)$, the density of the sum would

have generating function $\sum_k (p_1 p_{k-1} + p_2 p_{k-2} + \cdots) x^k = g^2(x)$. The mean of the random variable can be be computed by differentiating the generating function, $< X >= \sum_k k p_k = g'(1)$.

Another example more apropos to this paper is the following: Suppose we select a fraction α of the stubs[2] from a network whose degree distribution has generating function $g(x)$. Then what proportion of nodes will *not* be attached to any of the stubs we selected?

$$\sum_k p_k (1-\alpha)^k = g(1-\alpha)$$

Meanwhile the degree distribution of those not attached to a selected connection is generated by

$$\frac{g((1-\alpha)x)}{g(1-\alpha)}$$

We can do better by computing the explicit generating function for the joint degree distribution of selected and unselected stubs. This is accomplished by applying a second variable to the generating function. Let x correspond to selected stubs and y correspond to unselected stubs. The probability of a degree k node having m of its stubs selected is $\binom{k}{m}\alpha^m(1-\alpha)^{k-m}$. Then the generating function will be of the form

$$\begin{aligned} g(x,y;\alpha) &= \sum_k \sum_{m=0}^{k} p_k \binom{k}{m} \alpha^m (1-\alpha)^{k-m} x^m y^{k-m}/c \\ &= \sum_k p_k (\alpha x + (1-\alpha)y)^k / c = g(\alpha x + (1-\alpha)y)/c \end{aligned}$$

where $c = g(\alpha + \beta)$ is a normalizing constant. This example is important, as it underlies the methodology employed in this paper. The situation would be identical if infection had spontaneously spread among a fraction α of the stubs and we asked how many nodes remained uninfected.

[2]In network vernacular a *stub* is one end of a *connection* between two nodes.

We will use an indirect approach in that we will not track the evolution of susceptibles and infecteds directly, but rather the number of stubs which are attached to susceptibles and infecteds. When an infected node transmits infection along one of its connections, we say the corresponding connection is *occupied*. The variable T will be the number of stubs emanating from susceptible nodes which are not paired with an infected or refractory alte. The variable W will be the number of stubs emanating from infected nodes which have not yet become infected or refractory. We will treat the simple case of a constant force of infection and constant recovery rate. The quantities of interest in the model are as follows:

- $r :=$ Force of infection. The probability per unit time of infection traversing a network connection.
- $\mu :=$ Recovery rate. The probability per unit time that a connection to an infected will become refractory.
- $n :=$ The population size.
- $z :=$ The average degree in the network.
- $T :=$ The number of all network connections to susceptible nodes which have not become refractory.
- $W :=$ The number of all network connections to infected nodes which have neither become occupied nor become refractory.
- $\alpha :=$ The proportion of stubs not connected to an occupied or refractory stub, i.e. the survivor function of susceptible stubs.
- $\beta :=$ The proportion of stubs among susceptible nodes which are connected to refractory stubs.

- $S :=$ The number of susceptibles.
- $I :=$ The number of infecteds, including those in a refractory state.

It is important not to confuse stubs and connections. Two stubs are paired to form a connection. Stubs can be dormant, can be infected (infection has been transmitted by the stub to its alter), or can be refractory. In particular, it is possible for one stub to be refractory while its alter is infected. However if just one stub in a connection is infected, we say the corresponding connection is occupied. The dynamics proposed below do not keep track of the number of occupied connections, but rather of the number of stubs paired with infected or refractory alters. This is a pragmatic approach, as a susceptible can be defined as a node for which all of its stubs are not connected with infected alters.

During the course of an epidemic, a node may be connected to a refractory stub, an infected stub, or a dormant stub. The different types of connections can be factored into the generating function by using multiple variables. Let the variable x correspond to the number of stubs paired with dormant alters, and y correspond to the number of stubs paired with refractory alters. Note that at any given time, a susceptible will not have any stubs connected to an infected alter by definition. Since we are only interested in the degree distribution of susceptibles, we will not introduce a variable for the number of infected stubs.

For susceptibles, stubs will be distributed among refractory connections and unoccupied/non-refractory connections. As defined above, α is the probability of having the latter type of connection, while β is the probability of the former. The generating function for the degree distribution among susceptibles will be

$$\sum_k p_k(\alpha x + \beta y)^k / c = g(\alpha x + \beta y)/g(\alpha + \beta) \tag{4.2}$$

The quantity T is easy to derive by similar logic. The probability of a node having degree k and contributing m stubs to T is

$$p_k \binom{k}{m} \alpha^m \beta^{k-m}$$

So in terms of the PGF, the number of stubs emanating from susceptibles which do not have refractory alters will be

$$T = n \frac{d}{dx}[g(\alpha x + \beta y)]_{x=1,y=1} = n\ \alpha\ g'(\alpha + \beta) \tag{4.3}$$

α and β will change over the course of the epidemic, thereby controlling the evolution of the degree distribution (4.2). It remains to determine the dynamics of these parameters. At any given time, the hazard rate for an unoccupied stub being connected to an infected stub is rp_W, where $p_W = W/(W + T)$ is the proportion of non-refractory/unoccupied stubs connected to infecteds. Likewise, the hazard rate for becoming connected to a refractory stub is μp_W. Recall α is the survivor function for stubs not connected to occupied or refractory stubs; thus its dynamics is governed by

$$\dot{\alpha} = -\alpha(r + \mu)p_W \tag{4.4}$$

The evolution of β is more complicated. The probability of a stub connected to a susceptible node surviving to a time t is of course α. At time t, the hazard of connecting to a refractory stub is μp_W. Then we have the following:

$$\dot{\beta} = \alpha\ \mu\ p_W \tag{4.5}$$

The dynamics of W is dependent both on the outflow of stubs becoming occupied and refractory, plus the inflow of stubs from newly infected nodes. Note that the total degree mass of the network, $M = nz$ is conserved. If we denote by $\mathcal{X}$ the

stubs which are either occupied or refractory, we have the identity $W = M - T - \mathcal{X}$. Differentiating gives $\dot{W}$.

$$\dot{W} = -\dot{T} - \dot{\mathcal{X}} \tag{4.6}$$

$\dot{\mathcal{X}}$ is quite simple. When a network connection becomes occupied or refractory, the two stubs making up the connection change state. Then $\dot{\mathcal{X}}$ increases at twice the rate at which stubs from W become refractory or occupied.

$$\dot{\mathcal{X}} = 2(r + \mu)W$$

Differentiating equation (4.3) and using equation (4.4) gives

$$\dot{T} = -(r + \mu)p_W T - p_W \; r \; n \; \alpha^2 \; g''(\alpha + \beta) \tag{4.7}$$

Finally, combining equations (4.6), (4.1), and (4.7) we have

$$\dot{W} = (r + \mu)(p_W T - 2W) + p_W \; r \; n \; \alpha^2 g''(\alpha + \beta) \tag{4.8}$$

$$= p_W(r \; n \; \alpha^2 g''(\alpha + \beta) - (r + mu)(2W + T)) \tag{4.9}$$

This completes the model.

Once the model has been integrated the number of susceptibles can be determined by applying the PGF to distribution parameters α and β. At a given time t, the number of susceptibles S is

$$S = n \sum_k \sum_{m=0}^{k} p_k \binom{k}{m} \alpha^m \beta^{k-m} \tag{4.10}$$

$$= n \sum_k p_k (\alpha + \beta)^k = n \; g(\alpha + \beta) \tag{4.11}$$

The number of infecteds including those who have recovered is $I = n - S$.

4.2 Formal model specification

In addition to the previous intuitive discussion, the model is here placed on a more rigorous basis. The dynamical equations are derived starting from basic definitions and first principles.

4.2.1 Definitions

The undirected network can be defined as a graph $\mathcal{G} = \{V, \mathcal{E}\}$ consisting of a set of vertices V corresponding to the nodes in the network, and a set of edges $\mathcal{E}$ with elements of unordered pairs of vertices, $\{a, b\}$ where $a, b \in V$. We say that two vertices a, b are *neighbors* or *neighboring eachother* or simply *connected* if there exists an edge $e = \{a, b\} \in \mathcal{E}$.

At any point in time, a vertex can be classified as susceptible, infectious, or recovered. Let the disjoint sets S, J, R denote the set of vertices classified as susceptible, infectious, or recovered respectively.

As stated in the previous section, infectious vertices $a \in J$ will infect neighboring susceptible vertices $b \in S$ at a constant rate r. Infectious vertices $a \in J$ will become recovered (move to set R) at a constant rate μ.

Although the network is undirected in the sense that any two neighboring vertices can transmit infection to one another, we wish to keep track of who infects who. Therefore, for each edge $\{a, b\} \in \mathcal{E}$, let there be two arcs, which will be defined to be the ordered pairs (a, b) and (b, a). Let $\mathcal{A}$ denote the set of all arcs in the network. The first element in the ordered pair (a, b) will frequently be called the *ego* and the second element the *alter*.

Further define $\mathcal{A}_S$ as the set of arcs (a, b) such that the first element a belongs

to the set of susceptible vertices S. $\mathcal{A}_S = \{(a,b)|a \in S\}$. Similarly define the analagous sets $\mathcal{A}_J$ and $\mathcal{A}_R$ as the set of arcs such that the first element belongs to sets J or R respectively.

Also define the set $\mathcal{A}_O$ as the set of arcs such that the first element has transmitted infection to the second element. $\mathcal{A}_O = \{(a,b)|a \text{ transmits infection to } b\}$. In particular, note that this definition does not depend on the state of the second element b, which may already be infected or recovered.

For all of the arc-set definitions, let the sets A_{-X} denote the set of arcs such that the *second* element belongs in the corresponding set of vertices X. For example, the set $\mathcal{A}_{-O}$ will denote the set of arcs (a,b) such that the vertex b has transmitted infection to the vertex a: $\mathcal{A}_{-O} = \{(a,b)|b \text{ transmits infection to } a\}$.

4.2.2 Dynamics

A susceptible vertex $v \in S$ by definition is not part of any arc which has transmitted infection: $\mathcal{A}_S \cap \mathcal{A}_{-O} = \emptyset$. A susceptible vertex may however be connected to infectious vertices and vertices which were infected but have since become recovered. The set of arcs capable of transmitting infection to a susceptible vertex at a given time are those such that ego is a susceptible vertex and alter is not recovered. This set will be denoted by $\mathcal{A}_T = \mathcal{A}_S \setminus \mathcal{A}_{-R}$.

Similarly, $\mathcal{A}_W$ will denote the set of arcs which might transmit infection to a susceptible vertex. In other words, the ego is infectious and has not transmitted to the alter,and the alter is not already recovered and the alter has not already transmitted to ego. Formally, $\mathcal{A}_W = \mathcal{A}_J \setminus (\mathcal{A}_O \cup \mathcal{A}_{-O} \cup \mathcal{A}_{-R})$.

Let W, T be scalar quantities denoting the size of the sets $\mathcal{A}_W, \mathcal{A}_T$ respectively.

The sets of arcs $\mathcal{A}_W, \mathcal{A}_T$ along with the vertices the constitute them form a

closed subgraph of $\mathcal{G}$. If $(a,b) \in W$ then $a \in J$, $(a,b) \notin \mathcal{A}_O \cup \mathcal{A}_R$ and $(b,a) \notin \mathcal{A}_O \cup \mathcal{A}_R$. Then if $b \in J$ then $(b,a) \in \mathcal{A}_W$. If $b \in S$ then $(b,a) \in T$.

Because $\mathcal{A}_W \cup \mathcal{A}_T$ form a random subgraph, the probability that an alter of an arc in $\mathcal{A}_W \cup \mathcal{A}_T$ is in J is $W/(W+T)$.

In a time interval dt, an expected number of arcs $rWdt$ in $\mathcal{A}_W$ will transmit infection from ego to alter. An expected number μJdt infectious nodes will become recovered. Each infectious vertex which becomes recovered is an ego of on average W/J arcs in $\mathcal{A}_W$, so that the expected number of arcs from $\mathcal{A}_W$ that go to set $\mathcal{A}_R$ in time dt is $\mu Jdt(W/J) = \mu Wdt$.

Then in time dt, among arcs in $\mathcal{A}_T$, an expected number $(r+\mu)Tdt(W/(W+T))$ have an alter which transmits infection to ego or becomes recovered. The instantaneous probability (or hazard rate) of this event occuring in time dt will be denoted by

$$\lambda(t) = \lambda_1(t) + \lambda_2(t) \tag{4.12}$$

$$\tag{4.13}$$

where

$$\lambda_1(t) = rp_W \tag{4.14}$$

$$\lambda_2(t) = \mu p_W \tag{4.15}$$

and $p_W = W/(W+T)$.

Theorem. The probability of a degree $k = 1$ node $v \in S$ at time t is $\alpha + \beta$ where

$$\dot{\alpha} = -\alpha\lambda(t) \tag{4.16}$$

$$\dot{\beta} = \alpha\lambda_2(t) \tag{4.17}$$

Sketch of proof. Note that $\dot{\alpha} + \dot{\beta} = -r\alpha p_W$. Consider susceptible vertex $v \in S$. Case 1: $(v, a) \in \mathcal{A}_{-R}$. Then there is no hazard of v becoming infected at time t or at any time $\tau > t$. Case 2: $(v, a) \in \mathcal{A}_T$. Then the hazard that v becomes infected at time t is $\lambda_1(t)dt$. The hazard of a becoming refractory is $\lambda_2(t)$. Say α is the fraction of arcs such that neither event has occured at time $\tau < t$. Then

$$\frac{\dot{\alpha}}{\alpha} = -\lambda_1(t) - \lambda_2(t) \Rightarrow \tag{4.18}$$

$$\dot{\alpha} = -\alpha\lambda(t) \tag{4.19}$$

Say β is the fraction of arcs such that the alter is recovered. β grows at the rate that α surviving arcs have alter which becomes recovered, an event with probability $\lambda_2(t)$.

$$\dot{\beta} = \alpha\lambda_2(t) \tag{4.20}$$

Theorem. The probability of a degree k vertex surviving (remaining susceptible) to a time t is $(\alpha + \beta)^k$. Each arc independently experiences hazard λ_1 of transmitting infection. The probability that infection has not been transmitted on any arc connecting a degree k vertex is the product of the probabilities that it has not been transmitted on any one.

Corollary. The proportion of susceptibles at any time t is

$$g(\alpha + \beta) = \sum_k p_k(\alpha + \beta)^k \tag{4.21}$$

Corollary. The size of the set $\mathcal{A}_T$ at any time t is $n\alpha g'(\alpha + \beta)$. Note that an expected αk arcs connected to a degree k susceptible vertex will not have a

refractory alter. So

$$T = n\sum_k (k\alpha)p_k \text{Pr.}\{ \text{ degree k vertex is suceptible } \} \tag{4.22}$$

$$= n\alpha \sum_k k p_k (\alpha+\beta)^k \tag{4.23}$$

$$= n\alpha g'(\alpha+\beta) \tag{4.24}$$

Also note by applying product and chain rules we have

$$\dot{T} = n\dot{\alpha}g'(\alpha+\beta) - n\alpha g''(\alpha+\beta)(r\alpha p_W) \tag{4.25}$$

It remains to derive the dynamics for the number of arcs which can transmit infection, $\mathcal{A}_W$.

Let $\mathcal{A}_\chi = (\mathcal{A}_W \cup \mathcal{A}_T)^c$ be the arcs outside of sets $\mathcal{A}_W$ and $\mathcal{A}_T$. When an infectious vertex becomes recovered, all the arcs for which it is ego or alter leave the set $\mathcal{A}_W \cup \mathcal{A}_T$. Similarly, when an infectious vertex transmits infection along an arc (a, b), then both arcs (a, b) and (b, a) leave the set $\mathcal{A}_W \cup \mathcal{A}_T$. This event occurs at a rate proportional to the size of $\mathcal{A}_W$, so that

$$\dot{\chi} = 2(r+\mu)W \tag{4.26}$$

Note that $|\mathcal{A}| = T + W + \chi$ is constant. Then

$$\dot{W} = -\dot{T} - \dot{\chi} \tag{4.27}$$

4.3 Examples

The model has been tested on several common degree distributions:

- Poisson: $p_k = \frac{z^k e^{-z}}{k!}$. This is generated by

$$g(x) = e^{z(x-1)} \tag{4.28}$$

- Power-law. For our experiments, we utilize power-laws with exponential cutoffs κ: $p_k = \frac{k^{-\gamma}e^{-k/\kappa}}{Li_\gamma(e^{-1/\kappa})}, k \geq 1$ where $Li_n(x)$ is the nth polylogarithm of x. This is generated by

$$g(x) = Li_\gamma(xe^{-1/\kappa})/Li_\gamma(e^{-1/\kappa}) \tag{4.29}$$

- Exponential: $p_k = (1 - e^{-1/\lambda})e^{-\lambda k}$. This is generated by

$$g(x) = \frac{1 - e^{-1/\lambda}}{1 - xe^{-1/\lambda}} \tag{4.30}$$

If a single node is chosen at random from the population and infected, we can anticipate the following initial conditions: The survivor function for uninfected stubs, α, will begin at 1 and evolve downwards. β will begin at 0 and evolve upwards. T will be equivalent to the degree mass of the network minus the degree of the initial infected. And W will be the degree of the initial infected. We take the degree of the initial infected to be the average degree within the network. These are the initial conditions used in the trials shown in figure 4.1 and 4.2.

Figure 4.1 shows the disease incidence for each of the degree distributions (4.28), (4.29), and (4.30), with a force of infection $r = .2$ and mortality $\mu = .1$. The parameters of the degree distributions were chosen so that each network has an identical average degree of 3. That is, the density of connections in each network is the same. Nevertheless, there is widely different epidemic behavior due to the different degree distributions.

A sense for the different dynamical behaviors of each of the three networks can be gotten from figure 4.1. Consistent with previous research, the degree distribution has a great impact on the final size of the epidemic [18, 20]. More importantly, the three networks exhibit widely varying dynamical behavior. In particular, note

that the power law network experiences epidemics which accelerate very rapidly. Such epidemics enter the expansion phase virtually as soon as the first individual in the network is infected. Both the Poisson and exponential networks experience a lag before the expansion phase of the epidemic. These observations are consistent with the findings in [5] that the timescale of epidemics shortens with increasing contact heterogeneity. Pure power laws have an infinite second moment, and therefore have a minimally short time-scale. This has important implications for intervention strategies, as it is often the case that interventions are planned and implemented only after a pathogen has circulated in the population for some time. If an epidemic were to occur in the power-law network, there would be little time to react before the the infection had reached a large proportion of the population.

Several other variables of interest are computed as a byproduct of the model. Figure 4.2 shows the most important for the power law trial described above. α shows the proportion of stubs not connected to an occupied or refractory alter. β shows the proportion of stubs *among susceptibles* connected to a refractory alter. These variables do not quite move in tandem and may cross each other. Also shown is W (rescaled by population size n) which is similar to the hazard rate of becoming infected ($rW/(W+T)$). The epidemic ceases only when W reaches negligible levels.

Something offered by this model and not to the author's knowledge seen previously, is an explicit calculation for how the degree distribution of susceptibles evolves over the course of the epidemic. The infection will clearly tend to strike more highly connected individuals before more isolated individuals. Thus we expect the degree distribution to become bottom-heavy, as high degree nodes are gradually weeded out of the population. This is indeed observed in figure 4.3 for

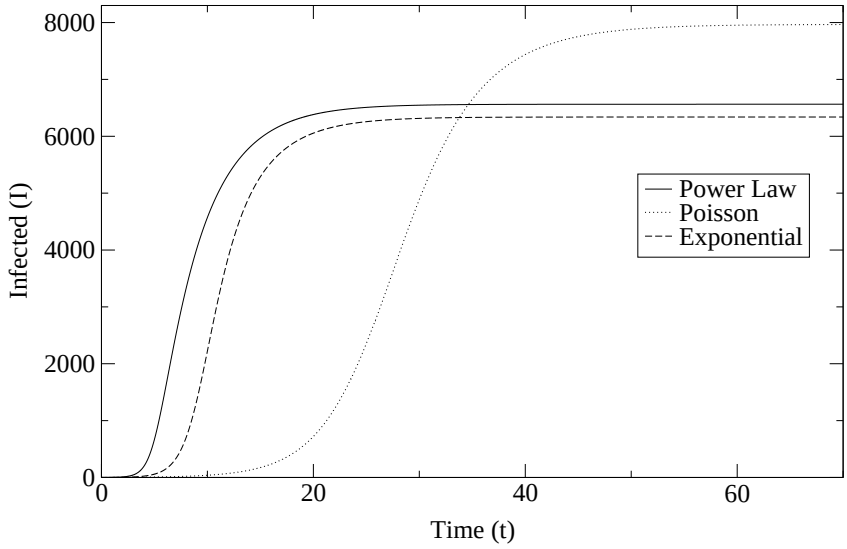


Figure 4.1: The number of infecteds (including recovered) is shown versus time for an SIR model on three networks. Force of infection and mortality are constant: $r = 0.2$, $\mu = 0.1$. The networks have Poisson ($z = 3$), power law ($\gamma = 1.615, \kappa = 20$), and exponential ($\lambda = 3.475$) degree distributions. Each of these degree distributions has an average degree of 3.

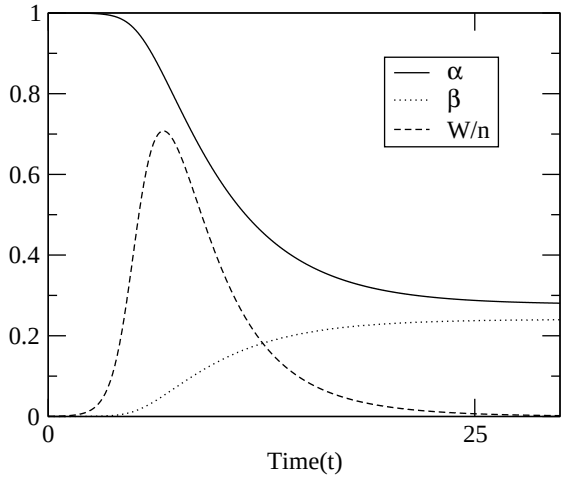


Figure 4.2: α, β, and W/n are shown versus t for a power law network with exponent $\kappa = 1.615$ and exponential cutoff $\kappa = 20$. Force of infection and mortality are constant: $r = 0.2$, $\mu = 0.1$.

the Poisson trial described above.

Recall that the degree distribution of susceptibles is generated by equation (4.2) and that we retrieve the explicit degree distribution by differentiation:

$$p_k = [(\frac{d^k}{dx^k} g(x)]_{x=0}/k! \tag{4.31}$$

Applying this to the Poisson PGF (equation (4.28)) gives

$$p_k = \frac{(z\alpha)^k e^{-z\alpha}}{k!} \tag{4.32}$$

We recognize this as simply the Poisson distribution with an adjusted parameter $z \times \alpha$.

Previous work [20] has shown that there is a critical transmissibility above which an epidemic will reach a fraction of the population in the limit as n goes to infinity. Below that threshold, the epidemic is limited to a finite-sized outbreak. Figure 4.4 shows the qualitatively different dynamical behavior of outbreaks below

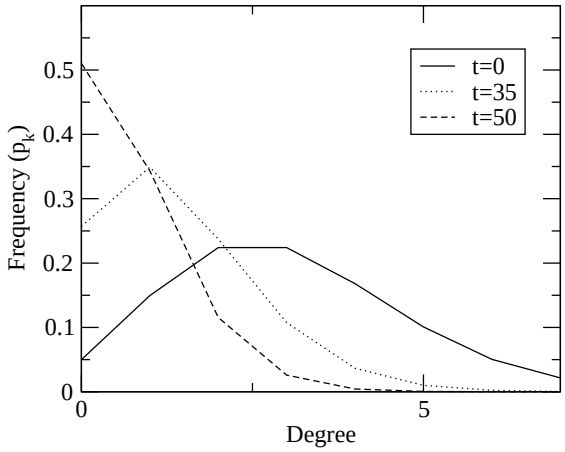


Figure 4.3: The degree distribution (equation (4.32)) for susceptibles is shown at three different times during the course of an epidemic on a Poisson network ($z = 3$). Force of infection and mortality are constant: $r = 0.2$, $\mu = 0.1$.

the phase transition for networks with a Poisson distribution. Note that these outbreak sizes are independent of the population size, n, in contrast to the incidence curves above the phase transition which are sensitive to n both in the time-scale of the epidemic and the number ultimately infected.

Define the *transmissibility*, τ, of the disease as the probability that the infection will traverse a network connection between and infected and a susceptible[3]. With constant force of infection and mortality

$$\tau = \frac{r}{r + \mu}$$

What is the critical transmissibility that defines the phase transition? Recall that the epidemic is complete when W is negligible and decreasing. If W is decreasing at $t = 0$ then the epidemic will necessarily die out without reaching a fraction of

[3] τ is related to to the traditional R_0 through the degree distribution. See [18]

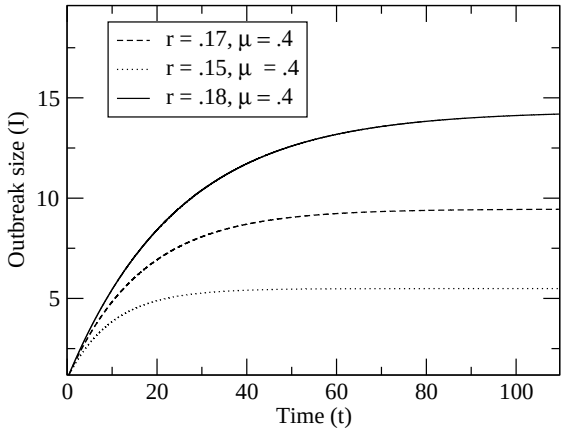


Figure 4.4: The number of infecteds (including recovered) is shown versus time for an SIR model on a Poisson network ($z = 3$). Each of these trials are below the critical level of transmissibility required to sustain an epidemic. Mortality is constant, $\mu = 0.4$, while three different levels of the force of infection are tried, $r = 0.15, 0.17, 0.18$.

the population. The critical point occurs where

$$\dot{W}_{t=0} = 0 = -\dot{T} - \dot{\mathcal{X}}$$

Applying equations (4.7) and (4.1)

$$\begin{aligned} 0 &= \frac{\alpha W}{W+T}[(r+\mu)g'(\alpha+\beta) + \alpha\ r\ g''(\alpha+\beta)] - 2(r+\mu) \\ \frac{r+\mu}{r}\Big(\frac{\alpha n}{W+T}g'(\alpha+\beta) - 2\Big) &= \frac{-\alpha^2\ n\ g''(\alpha+\beta)}{W+T} \\ \frac{r}{r+\mu} = \tau &= \frac{2W+T}{n\ \alpha^2 g''(\alpha+\beta)} \end{aligned}$$

At $t = 0$, $\alpha = 1$, $\beta = 0$, $W \approx 0$ and $T \approx n\ g'(1)$. Then

$$\tau^* = g'(1)/g''(1) \tag{4.33}$$

This is in agreement with previous results based on bond-percolation theory [20].

4.4 Discussion

The statistical properties of SIR epidemics in random networks have been understood for some time, but the explicit dynamics have been understood mainly through simulation. This paper has addressed this shortcoming by proposing a system of differential equations to model SIR in random networks.

It should be noted that the SI dynamics are a special case of this model ($\mu = 0$), in which case the ultimate extent of the epidemic is simply the giant component [19][4] of the network.

The distribution of contacts, even holding the density of contacts constant, has enormous impact on epidemic behavior. This goes beyond merely the extent of the epidemic, but as shown here, the dynamical behavior of the epidemic. In particular, the distribution of contacts plays a key role in determining the onset of the expansion phase.

The distribution dynamics from equation (4.2) and shown in figure 4.3 have important implications for vaccination strategies. Previous work [16, 14] has focused on determining the critical levels of vaccination required to halt or prevent an epidemic. It is usually taken for granted that contact patterns among susceptibles are constant. Furthermore, most widespread vaccinations occur only once an epidemic is underway. Future research could be enhanced by considering optimal vaccination levels when the epidemic proceeds unhindered for variable amounts of time.

It is hoped that the distribution dynamics described in this paper will find applications beyond modeling heterogeneous connectivity. The dynamic PGF approach

[4]The *giant component* of a network, if it exists, is the largest set of nodes such there exists a path between any two of them; furthermore the giant component must occupy a fraction of the network in the limit as network size goes to infinity.

may be used to capture other forms of heterogeneity, such as of susceptibility, mortality, and infectiousness.

REFERENCES

[1] M. Altmann. Susceptible-infected-removed epidemic models with dynamic partnerships. *J. Math Biol.*, 33:661–75, 1995.

[2] R.M. Anderson and R.M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, 1991.

[3] H. Andersson. Limit theorems for a random graph epidemic model. *Ann. Appl. Probab.*, 8:1331–1349, 1998.

[4] K.B. Athreya and P. Ney. *Branching Processes*. Springer, New York, 1972.

[5] M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *J. of Theor. Biol.*, 235:275–288, 2005.

[6] N. Becker. Estimation for discrete time branching processes with applications to epidemics. *Biometrics*, 33:515–522, 1977.

[7] M. Boguna, R. Pastor-Satorras, and A. Vespignani. Epidemic spreading in complex networks with degree correlations. In J.M. Rubi et. al., editor, *Statistical Mechanics of Complex Networks*, Berlin, 2003. Springer Verlag.

[8] Z. Dezso and A.L. Barabasi. Halting viruses in scale-free networks. *Phys. Rev. E*, 65:055103(R), 2002.

[9] O. Diekmann and J.A.P. Heesterbeek. *Mathematical epidemiology of infectious diseases. Model building, analysis and interpretation*. John Wiley & Sons, Ltd., Chichester, 2000.

[10] T.D. Eames and M.J. Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *PNAS*, 99:13330–13335, 2002.

[11] S. Eubank, H. Guclu, V.S. Anil-Kunar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic social networks. *Nature*, 429:180–184, 2005.

[12] C. Farrington, M. Kanaan, and N. Gay. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4:279–295, 2003.

[13] S. Gupta, R.M. Anderson, and R.M. May. Networks of sexual contacts: Implications for the pattern of spread of hiv. *AIDS*, 3:807–817, 1989.

[14] M.E. Halloran, I. Longini, A. Nizam, and Y. Yang. Containging bioterrorist smallpox. *Science*, 298:1428, 2005.

[15] T.E. Harris. *The Theory of Branching Processes*. Springer, Berlin, 1963.

[16] E.H. Kaplan, D.L. Craft, and L.M. Wein. Emergency response to a smallpox attach: The case for mass vaccination. *PNAS U.S.A.*, 99:10935, 2002.

[17] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, and Y. Aberg. The web of human sexual contacts. *Nature*, 411:907–908, 2001.

[18] L.A. Meyers, B. Pourbohloul, M.E.J. Newman, D.M. Skowronski, and R.C. Brun-ham. Network theory and sars: Predicting outbreak diversity. *J. Theor. Biol.*, 232:71–81, 2005.

[19] M. Molloy and B. Reed. The size of the giant component of a random graph

with a given degree sequence. *Combinatorics, Probability and Computing*, 7:295–305, 1998.

[20] M.E.J. Newman. The spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, 2002.

[21] M.E.J. Newman, D.J. Watts, and S.H. Strogatz. Random graph models of social networks. *PNAS USA*, 99:2566–2572, 2002.

[22] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, 2001b.

[23] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E*, 63:066117, 2001c.

[24] R. Pastor-Satorras and A. Vespignani. *Handbook of Graphs and Networks: From the Genome to the Internet*, chapter Epidemics and immunization in scale-free networks. Wiley-VCH, Berlin, 2002.

[25] J. Saramki and K. Kaski. Modelling development of epidemics with dynamic small-world networks. *J. Theor. Biol.*, 234:413–421, 2005.

[26] S.H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[27] V.M. Veliov. On the effect of population heterogeneity on dynamics of epidemic diseases. *J. Math. Biol.*, 51:123–143, 2005.

[28] C.P. Warren, L.M. Sander, I. Sokolov, C. Simon, and J. Koopman. Percolation on disordered networks as a model for epidemics. *Math. Biosci.*, 180:293–305, 2002.

[29] H.S. Wilf. *Generatingfunctionology.* Acedmic Press, Boston, 1994.

CHAPTER 5

PROBABILITY BASED ESTIMATION THEORY FOR RESPONDENT DRIVEN SAMPLING

Chain-referral sampling has emerged as a powerful method for sampling hard-to-reach or hidden populations. Such sampling methods are favored for such populations because they do not require the specification of a sample frame.

The lack of a sampling frame means that the survey data from a chain-referral sample is contingent on a number of factors outside the researcher's control such as the social network on which recruitment takes place. The major challenge of chain-referral sampling has been to understand an unconventional sampling process and to base estimates on the resulting data. In this article we draw on previous research on Respondent Driven Sampling (RDS) [5, 6, 12] to show that with a few plausible assumptions about the recruitment process and the social network, it is possible to specify selection probabilities for individuals in the target population and to apply traditional probability theory to the problem of statistical inference.

The new estimator we present here is similar to estimators originally proposed in the RDS literature [5, 6, 12], although the new estimator is based on a different theoretical foundation. The classical RDS estimator is based largely on Markov chain theory and social network theory. Our new estimator relies on Markov chain sampling theory [4, 9] and the theory of sampling with unequal probabilities [3, 1].

This paper should be viewed as part of an established and growing literature on network sampling [13, 14]. Birnbaum and Sirken (1965) were the first to consider sampling in affiliation networks, such as the networks of patients and health-care providers. Felix-Medina and Thompson (2004), Spreen and Zwaagstra (1994), and Rothenberg et al. (1995) have considered network sampling for hidden populations

using link-tracing [1] or snowball[2]designs. Frank (1978) has considered the problem of estimating topological features of social networks given a standard random sample from a network. Work by Frank and Snijders (1994) and recently Thompson (1998) has focused on deriving unbiased estimates from snowball-type and link-tracing samples, and in this sense is most similar to this work.

In section 5.1 we review the basics RDS methodology and why it is favored over other chain-referral methods. In section 5.2 we introduce a new RDS estimator which offers several advantages over the traditional methods of RDS estimation. Section 5.3 contains an analytical comparison of the new estimator to classical RDS methodology. Section 5.4 contains a prospective variance estimator, and finally section 5.5 presents the results of a simulation study to compare the new and old estimators.

5.1 Respondent Driven Sampling

Respondent Driven Sampling (RDS) is a rigorous system of chain-referral sampling which allows for statistical inference of the target population by controlling for the sources of bias usually associated with chain-referral sampling.

RDS is now being implemented in the US and around the world to study hard to reach or "hidden" populations. The Centers for Disease Control and Prevention has announced that it will use RDS to track HIV-risk behavior among injectors in 25 cities in the US, and Family Health International, the largest non-profit in global public health, is using it in more than a dozen countries [8, 6]. The main

[1]Link tracing designs combine traditional cluster sampling or standard random sampling with chain-referral methods.

[2]*Snowball* usually refers to chain-referral designs which exhaustively map out social networks. This should be contrasted with the random-walk design considered in the present manuscript.

advantage of RDS is that it does not require an ordinary sampling frame. Thus it is effective for stigmatized, hidden, or hard to reach populations, for which the researcher lacks organizational or institutional access.

Chain-referral sampling data differs from ordinary samples in that the respondents are linked together by a chain of recruitments. In general, each respondent will have attributed to them a coupon with a unique serial number which was given to them by another respondent. They will also have a a limited number of coupons which they may give to other respondents. Thus it is possible to keep track of who recruited whom. Figure 5.1 shows an actual recruitment chain drawn from a RDS study of New York City jazz musicians [7].

RDS begins with the selection of an initial respondent, or "seed". Selection of the seed is typically non-random, such as via public venues or health centers. The seed is given a number of coupons to distribute to friends and acquaintances which can be redeemed by being interviewed. When interviewed, the new respondent is in turn given coupons to distribute, thereby perpetuating the sample chain.

Additionally, RDS requires that we keep track of the degree of each respondent. The degree of a node in a network is the number of connections to that node, i.e. the number of neighbors of that node. In the context of chain-referral sampling, the degree of an individual will be defined as the number of people that that person *could*, in principle, recruit. We consider undirected networks only, such that recruitment can take place in both directions across a social network connection.

We will assume that our chain-referral samples are with-replacement, that is, any individual may be recruited into the sample more than once. In practice, the condition of with-replacement sampling is rarely met. It is possible that participation in the study may alter the acceptance rate of individuals to participate

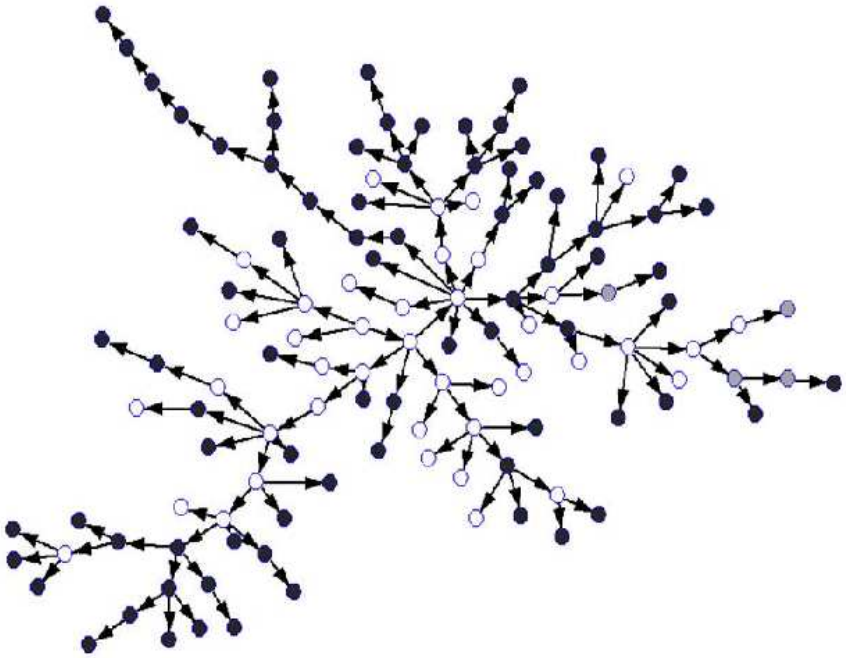


Figure 5.1: Example of a recruitment chain. This recruitment chain comes from a RDS study of jazz musicians in New York City [7]. Arrows indicate the direction of recruitment. The colors indicate the gender of each respondent: Black = Male, White = Female, Grey = Missing Data

Table 5.1: Notation used throughout this paper.

U is the set of all individuals in the population
S is the set of all individuals in the sample
$A, B, \ldots$ are disjoint sets of individuals
N_X is the number of elements in a set X
n_X is the number of sample units from set X
$P_A, P_B, \ldots$ are the population proportions of each type, A, B, etc.
$\vec{P}$ is the vector with elements P_A, P_B, etc.
R_{AB} is the number of recruitments from group A to group B
R_A is the total number of times people of type A are recruited
$\bar{R_A}$ is the total number of recruitments from people of type A
σ_{AB} is the probability of someone from set A selecting someone from set B
δ_i is the degree of individual i
δ_X Average degree of individuals from set X

in the study again. This could be a strong confounding factor if sampling with-replacement was allowed. But if the sampling fraction is very small, we can safely use results based on sampling with-replacement to the case of sampling without-replacement.

In the following treatment, we assume that each respondent recruits only one neighbor, although methods have been devised to compensate for the case where respondents may recruit more than one neighbor. Details on this method, called *demographic adjustment*, can be found in section 5.3.

Further, we assume that the sampling fraction is small, such that we can apply solutions for the sampling-with-replacement case. Refer to table 5.1 for a list of notation used throughout this article.

In developing our theory, we will rely on the following assumptions in addition to those mentioned above:

1. *Degree.* Respondents accurately report their degree in the network.

2. *Recruitment is random.* When recruiting others, respondents select uniformly at random from their personal network.

3. *Reciprocity.* Network connections are reciprocal. Respondents recruit those with whom they have a pre-existing relationship, such as acquaintances, friends, and those closer than friends. Such connections are reciprocal, e.g., my friends and acquaintances consider me to be a friend or acquaintance. Consequently, in network theoretic terms, the potential recruitment network is undirected, so if respondent a can recruit b, then b can also recruit a. This is required by the *reciprocity model* (Heckathorn 2002,Salganik and Heckathorn 2004) upon which the original RDS estimator is based. This is formally known as the *reciprocity hypothesis*.

4. *Convergence.* Recruitment is modeled as a Markov process (MP), where the state of the MP is the last individual recruited. Transition probabilities are described in section 5.2. We assume that the MP is irreducible and that each state has a finite return time. Therefore, a unique equilibrium to the MP exists and recruitment rapidly converges to this equilibrium. The implication is that after a modest number of steps, the sample composition becomes independent of the initial respondents ("seeds") who initiated the chain-referral process.

The irreducibility condition is equivalent to the condition that the social network is well-connected; that is to say, every node can be reached by a finite path from

any other node. Furthermore, our social networks are assumed to be finite (though very large), so the expected return time must be finite as well.

On the surface, the irreducibility assumption may seem unrealistic, especially for large populations, where it is most likely that some units will be isolated from the network as a whole. This is true, however it is usually not a cause for concern. It is known from random network theory that most networks possess a so-called *giant-component*, a subset of nodes such that a network path exists between any two and which occupies a non-vanishing fraction of the network as population size goes to infinity. The giant component usually encompasses the vast majority of the population, so long as some basic conditions are met. For instance, in pure random graphs, the giant component will consist of 99% of the population if nodes have just 5 links on average. RDS studies have typically exceeded this margin comfortably. In a study of NYC jazz musicians, respondents were found to have an average degree of 109 [7], while in a study of gay Latinos, respondents in San Francisco had an average degree of 8 and in Chicago had an average degree of 13 [11]. With that said, field RDS studies should come with the caveat that statistical inference is limited to the giant component, rather than the total population. But provided the giant component is very large, this is usually a minor distinction.

Furthermore, research on the *small-world* problem [15] has led to the observation that almost all social networks have very short mean path length. Consequently, there are relatively few intermediaries between any two randomly selected individuals in most social networks. In pure random networks [2, 10], path length grows logarithmically with population size. A consequence of this, is that the selection probability for any individual in the network will stabilize after just a few recruitments. In other words, the process will have no "memory" of past recruits.

Another assumption commonly called into question is that respondents recruit uniformly at random from their network neighbors. Indeed, it is difficult or impossible to enforce random recruitment among respondents, and in many cases, respondents may have special reasons for selecting a particular recruit. However, non-random recruitment, if it occurs, will not necessarily bias our estimator. As long as recruitment is not correlated with any variable important for estimation (e.g. the study-variable or degree), the aggregate effect is for recruitment to appear uniform-random.

Non-random recruitment would most obviously be evidenced by skewed and non-symmetric recruitment matrices. If, for instance, respondents preferred recruiting someone of type A, we would expect recruitment matrices with much more weight on elements R_{XA} than on elements R_{AX}. In fact, this is rarely observed. By now, strong empirical evidence [6] has built up that random recruitment holds in most cases. It is nevertheless a potential source of bias that practitioners should watch out for.

5.2 New estimators for Respondent Driven Sampling

It is often the case that it is more convenient to sample from a distribution other than the one we wish to use for estimation. In this case, the theory of Markov chain sampling has developed in order to sample from arbitrary distributions. The premise is to devise a Markov process (MP) such that the equilibrium distribution of the MP is identical to the distribution one wishes to sample from. It has further been shown that estimators based on a Markov chain samples are asymptotically

unbiased[3]. [4]

In contrast to traditional Markov chain sampling, we are not at liberty to devise the transition probabilities between our sampling units due to the lack of a standard sampling frame. Rather the transition probabilities are imposed on us by the nature of the chain referral sample and the properties of the social network. Nevertheless, the chain-referral sample will constitute a Markov chain which fits the criteria necessary to apply our theory.

In mathematical terms, a chain-referral sample is analogous to a random walk on a network. It has been shown [12] that a random walk on a network is a MP, which in equilibrium occupies a node with probability proportional to degree. We can then infer that a chain-referral sample will select individuals in the population with probability proportional to degree.

Let $\mathscr{E}$ be the incidence matrix of the network. $\mathscr{E}$ will have elements e_{ij} where $e_{ij} = 1$ if nodes i and j are connected, and will equal zero otherwise. Note that the degree of node i, δ_i, is the i'th row sum of $\mathscr{E}$, $\sum_j e_{ij}$. If the random walk is at node i at step t, the probability of node i choosing node j is $1/\delta_i = 1/\sum_j e_{ij}$. Denote this transition probability $\sigma_{ij}^{\mathscr{E}}$, and let the matrix with these transition probabilities be called $\sigma^{\mathscr{E}}$. The random walk on the network can therefore be considered a MP with transition probabilities $\sigma^{\mathscr{E}}$.

The random walks we consider are irreducible and finite, so there must be a unique equilibrium to this MP. Furthermore the MP will converge to this equilibrium. Consider the state vector x^* with elements

$$x_i^* = \delta_i / \sum_j \delta_j \tag{5.1}$$

[3]By *asymptotically-unbiased* we mean that any bias will be of the order $1/n$. Therefore, for meaningful sample sizes, any bias will be negligible.

It may be verified that x^* is an equilibrium to the MP given by $\sigma^{\mathscr{E}}$, and by our hypotheses, must also be a unique attracting equilibrium. Now that we have established that a chain-referral sample of the RDS type is a Markov chain sample, we may proceed to develop estimators for our target population. Using only the fact that RDS samples individuals with probability proportional to degree, we can develop a Hurwitz-Hansen (HH) type estimator for $\vec{P}$ [3, 1]. The derivation presented here uses a similar argument to that presented in [12] to estimate the average degree in a social network from chain-referral data.

HH estimators require knowledge of the selection probabilities, p_i, the probability that individual i will be selected at any stage of the chain-referral sample. Using the equilibrium condition (5.1), the selection probabilities will be

$$p_i = \frac{\delta_i}{N\delta_U} \tag{5.2}$$

which we can estimate as

$$\hat{p_i} = \frac{\delta_i}{N\hat{\delta_U}} \tag{5.3}$$

where $\hat{\delta_U}$ is the estimate of the average degree of the total population.

The $\hat{\delta_X}$ are easy to estimate. As in [12], we note that the average degree can be estimated as a ratio estimator of HH estimators.

$$\hat{\delta_U} = \frac{\sum_S \frac{\delta_i}{np_i}}{\sum_S \frac{1}{np_i}} = \frac{n}{\sum_S \delta_i^{-1}} \tag{5.4}$$

And for just one group, e.g. the subset A within the population

$$\hat{\delta_A} = \frac{n_A}{\sum_{A\cap S} \delta_i^{-1}} \tag{5.5}$$

This is the well-known formula for the harmonic mean, the mean of a quantity which is being sampled with probability proportional to its size.

Now let the variable y_i be some real-valued variable of interest. Let T_y represent the total value of y in the population, $\sum_U y_i$. y_i may represent continuous variables such as age or income, or dichotomous variables such as HIV status.

The HH estimator of the total y in the population, $\hat{T}_y$ is:

$$\begin{aligned}\hat{T}_y &= \tfrac{1}{n}\textstyle\sum_S \frac{y_i}{\hat{p}_i} \\ &= \tfrac{1}{n}\textstyle\sum_{i\in S} \frac{\hat{\delta_U} N y_i}{\delta_i} \\ &= \tfrac{\hat{\delta_U} N}{n}\textstyle\sum_S \delta_i^{-1} y_i\end{aligned}$$

If N is unknown, as is generally the case, we can still estimate the mean value of y as

$$< \hat{y} > = \frac{\hat{\delta_U}}{n} \sum_S \delta_i^{-1} y_i \tag{5.6}$$

Substituting the definition of $\hat{\delta_U}$ (eqn. 5.5), we arrive at the simple equation:

$$< \hat{y} > = \sum_{i\in S} \delta_i^{-1} y_i / \sum_{i\in S} \delta_i^{-1} \tag{5.7}$$

We will refer to this estimator as *RDS II* to distinguish it from the RDS estimator presented in section 5.3. Essentially equation 5.7 weights each case by the reciprocal of the corresponding degree value.

Suppose we are interested in estimating P_A, the proportion of the population of type A. Let y_i be the indicator function $I_A(i)$, which takes the value 1 if $i \in A$ and 0 otherwise. Using equation 5.7 we have

$$\hat{P_A} = \frac{\sum_{i\in A\cap S} \delta_i^{-1}}{\sum_{i\in S} \delta_i^{-1}} \tag{5.8}$$

There is an alternative form of equation 5.8 worth mentioning, as it gives some intuition for how our estimator works. With a little manipulation we get

$$\hat{P_A} = \left(\frac{n_A}{n}\right)\left(\frac{\hat{\delta_U}}{\hat{\delta_A}}\right) \tag{5.9}$$

The first part of equation 5.9, $\left(\frac{N_A}{N}\right)$ is the proportion of the sample of type A. If our sample were a standard random sample this would be our estimate for P_A. The second part, $\left(\frac{\delta_U}{\delta_A}\right)$ expresses the correction due to network effects. For example, if $\delta_U > \delta_A$ we are under-sampling individuals of type A, and consequently we inflate our estimate.

Note that the initial recruits in a chain-referral sample (i.e. the "seeds") will generally be chosen non-randomly. It is usually prudent to exclude them from the estimator(eqn. 5.8), as well as the estimation of average degree (eqn. 5.5), though the estimator will be asymptotically unbiased even if they are included. The rationale for eliminating seeds is the same as that for using a "burn in" period during a MCMC sample. Any potential bias accruing from the initial seed selection will be lessened. The recruitments made by seeds are usually included, however, in the recruitment matrix (sec. 5.3 below). Experimental evidence for how long a burn-in period is best is currently lacking.

5.3 The classical RDS estimation procedure and its relation to RDS II

In [5, 6, 12], it was shown how to convert a chain-referral sample into a probability sample of individuals in the population and to produce unbiased estimates from chain-referral sample data. The original RDS estimator accounted for all of the sources of bias usually associated with chain-referral samples, such as oversampling well-connected individuals and non-random mixing in the population. Here we present a brief review of this methodology with the objective of elucidating the relationship between the new estimator (5.8) and the original estimator proposed

in [5, 6, 12].

The classical RDS estimator (henceforth alternately referred to as *RDS I*) relies on the theory of network balance between sub-groups in the population. The mass of network connections to and from every group can be estimated up to a constant factor. This gives us a system of balance equations for every pair of groups, which in turn can be used to solve for the relative size of each group.

Specifically, it was observed that $R_{AB}/\bar{R_A}$ is an unbiased estimate of σ_{AB}, the probability of someone of type A recruiting someone of type B. Furthermore, the connections from group A to group B must be equal to those from B to A by the reciprocity hypothesis. The number of connections from group A to group B will be proportional to $\sigma_{AB}P_A\delta_A$. Given n groups, this then leads to a system of $\binom{n}{2}$ balance equations:

$$\begin{aligned}
\sigma_{12}P_1\delta_1 &= \sigma_{21}P_2\delta_2 && (5.10)\\
\sigma_{13}P_1\delta_1 &= \sigma_{31}P_3\delta_3 && (5.11)\\
&\vdots && (5.12)\\
\sigma_{23}P_2\delta_2 &= \sigma_{32}P_3\delta_3 && (5.13)\\
&\vdots && (5.14)\\
\sigma_{(n-1)n}P_{n-1}\delta_{n-1} &= \sigma_{n(n-1)}P_n\delta_n && (5.15)
\end{aligned}$$

This system of equations can be used to solve for $\hat{\vec{P}}$, our estimate for the population proportion of each group. Of course, we must also normalize our solution by using $\sum_x P_x = 1$. This system of equations is over-determined for systems with more than two groups, such that least squares regression may be used to solve for $\hat{\vec{P}}$.

Two enhancements to *RDS I* were proposed in [6], which dramatically improve the precision of the estimator. The first considered adjustments for sample data

in which respondents could recruit more than one network-neighbor. In this case, it is possible that some groups in the population may systematically recruit more than other groups in the population, a phenomenon called *differential recruitment*, which can dramatically alter the composition of the sample. However, under the assumption that such data still provides us with an unbiased estimate of the group-to-group transition probabilities, σ_{AB}, we can deduce what the sample composition would be in the absence of differential recruitment.

Given the matrix of transition probabilities σ with elements σ_{AB}, the theoretical equilibrium sample composition is the vector x^* which satisfies

$$x^* = x^*\sigma \tag{5.16}$$

With a theoretical equilibrium sample distribution x^* and unbiased estimates of σ, we can postulate what form the recruitment matrix would take in the absence of differential recruitment. We call this matrix $\tilde{R}$, which will have elements proportional to the theoretical equilibrium composition times the unbiased transition probabilities.

$$\tilde{R}_{AB} = (n\ x^*{}_A)\ \sigma_{AB} \tag{5.17}$$

In other words, $\tilde{R}_{AB}$ is the theoretical number of recruitments from group A to group B in the absence of differential recruitment.

At this point, we may find that for some pair of groups, $\tilde{R_{AB}} \neq \tilde{R_{BA}}$. Because RDS randomly samples connections in the network, and the number of connections between any two groups must be identical, $\tilde{R_{AB}}$ and $\tilde{R_{BA}}$ will be two point estimates for the same quantity. Therefore, a more accurate estimate can be gained by averaging the values. Averaging over all pairs of groups gives us a symmetric recruitment matrix, which we call the *data-smoothed* recruitment matrix, $\tilde{R}_{DS}$

Transition probabilities can be re-computed from the data-smoothed recruitment matrix. Furthermore, if we assume that differential recruitment does not bias the estimated average group degrees, δ_X, then the RDS estimator can be re-calculated. We will refer to this estimator as *RDS I/DS*. Simulation studies have revealed that this estimator has markedly different properties from RDS I, most notably similar accuracy and increased precision. To base estimates on the data-smoothed recruitment matrix, we require that neither σ nor the estimated δ_X are biased by differential recruitment. This may not always be the case, but in practice has proven a reliable assumption. RDS I/DS estimates are also much closer, in general, to RDS II estimates.

This review of traditional RDS theory is pertinent, as RDS II is closely related to the classical RDS estimator, RDS I. Note that these similarities only exist when considering categorical variables, as RDS I is not adaptable to the estimation of continuous quantities.

Whenever the recruitment matrix is symmetric (that is, whenever $R_{AB} = R_{BA}\ \forall A, B$) the RDS I and RDS II estimators will coincide. Consequently, basing RDS estimates on the demographically adjusted and data-smoothed recruitment matrix will equalize these estimators.

To put this on firmer ground, let's collect all terms in P_A in the RDS I system of equations 5.10. For any group X,

$$P_X = \frac{P_A \delta_A \bar{R_X} R_{AX}}{\delta_X \bar{R_A} R_{XA}}$$

from which it follows that

$$\sum_X P_X = 1 = P_A \frac{\delta_A}{\bar{R_A}} \sum_X \frac{R_{AX} \bar{R_X}}{R_{XA} \delta_X}$$

Neglecting initial respondents (seeds), $\bar{R_X} = n_X$. That is to say, the number

of individuals of type X recruited into the study is the same as the number of individuals in the study of type X. Then solving for P_A we have

$$P_A = \frac{n_A}{\delta_A} \left(\sum_X \frac{R_{AX} n_X}{R_{XA} \delta_X} \right)^{-1} \tag{5.18}$$

Note that if $R_{AX} = R_{XA}$, then their ratio falls out of the equation. This is exactly what happens with the demographically adjusted and data smoothed recruitment matrix. Now observe that

$$\sum_X \frac{n_X}{\delta_X} = \sum_X \sum_{i \in S \cap X} \delta_i^{-1} = \frac{n}{\delta_U}$$

Substituting this into equation 5.18 yields equation 5.9, our estimator for RDS II. Thus, provided that $R_{YX} = R_{XY}$ for all groups X and Y, these two estimators will coincide.

In passing, note that a parsimonious way of expressing the demographically adjusted RDS II estimator is the following

$$\hat{P_A} = x_A^* \left(\frac{\hat{\delta_U}}{\hat{\delta_A}} \right) \tag{5.19}$$

Referring back to equation 5.9, it is clear that the only part of the RDS II estimator that will be biased is the sample proportion of type A, n_A/n. To correct for the bias, simply substitute the equilibrium composition x_A^* for the sample proportion of type A.

5.4 Variance estimation

The complicated design of RDS creates numerous challenges for variance estimation. It is tempting to apply the well known variance estimator for HH estimators [3]

$$\hat{V}_{HH}(\hat{T}_y) = \frac{1}{N^2 n(n-1)} \sum_S \left(\frac{y_i}{p_i} - \hat{T}_y \right)^2 \tag{5.20}$$

which when estimating the average value of y, becomes

$$\hat{V}_{HH}(<\hat{y}>) = \frac{1}{n(n-1)} \sum_{S} \left(\frac{y_i \hat{\delta_U}}{\delta_i} - <\hat{y}> \right)^2 \tag{5.21}$$

In fact, outside of a few special cases, this variance estimator performs quite poorly. The reason is that there are multiple sources of variance in the estimator (5.8). Sampling with non-identical selection probabilities is considered in the above variance estimator. But additionally, an RDS sample constitutes an MCMC sample of the social network, with transition probabilities σ. Therefore, sample units will be correlated, and in general, it is necessary to take this correlation into account when estimating variance. Below, we derive such a variance estimator. We conclude with the estimation of variance for $\hat{P_A}$ expressed in the following the equation:

$$\hat{V}_{P_A} = \hat{V}_{HH} + \frac{\hat{P_A}^2}{n} \left((1-n) + \frac{2}{n_A} \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\sigma^{i-j})_{AA} \right) \tag{5.22}$$

where

$$\hat{V}_{HH} = \hat{V}(Z_i)/n = \frac{1}{n(n-1)} \sum_{S} \left(Z_i - \hat{P_A} \right)^2 \tag{5.23}$$

and

$$Z_i = \hat{\delta_U} \delta_i^{-1} I_A(i) \tag{5.24}$$

The derivation that follows is based on the estimation of P_A, the proportion of the population of type A.

A little rearrangement shows the form of the estimator (5.8) can be expressed as

$$<\hat{y}> = \frac{1}{n} \sum_{S} Z_i =< Z > \tag{5.25}$$

An unbiased estimate of variance for the Z_i taken individually is

$$\hat{V}(Z_i) = \frac{1}{n-1} \sum_{S} \left(Z_i - \hat{P_A} \right)^2 \tag{5.26}$$

and, $\hat{V}_{HH} = \hat{V}(\frac{1}{n}\sum_S Z_i)$ is the naive estimate of variance for $< \hat{Z} >$ if we assumed each Z_i were i.i.d. random variables. In fact, there will frequently be non-trivial covariance between sample units determined by their proximity in the recruitment chain. As in section 5.3 we will use the transition probabilities σ to describe the probability of someone of type X recruiting someone of type Y. Note that this is a simplification; it is not always the case that recruitment can be modelled as a first order Markov process with transition probabilities σ, and the node-specific transition probabilities $\sigma^{\mathscr{E}}$ are almost always unknown.

We wish to find the variance

$$\begin{aligned}
&\hat{V}(Z_1 + Z_2 + \cdots + Z_n) = \\
&\hat{V}(Z_1 + \cdots + Z_n - 1) + 2\ cov(Z_1 + \cdots + Z_{n-1}, Z_n) + \hat{V}(Z_n) \\
&= \hat{V}(Z_1 + \cdots + Z_{n-2}) + 2\ cov(Z_1 + \cdots + Z_{n-2}, Z_{n-1}) + \\
&2\ cov(Z_1 + \cdots + Z_{n-1}, Z_n) + \hat{V}(Z_{n-1}) + \hat{V}(Z_n) \\
&\vdots
\end{aligned}$$

where in general

$$\begin{aligned}
&cov(Z_1 + \cdots + Z_{m-1}, Z_m) = \\
&E\ (Z_1 + \cdots + Z_{m-1} - (m-1) < Z >)(Z_m - < Z >) \\
&= -(m-1) < Z >^2 + \sum_{i=1}^{m-1} E\ Z_i Z_m
\end{aligned}$$

In the above equation, the expected value of the product $Z_i Z_m$ must be computed. In this case the correlation between these sample elements becomes important. Given that unit i is of type A (such that $y_i = 1$), the probability of unit m also

being of type A is $(\sigma^{m-i})_{AA}$, and thus

$$E\ Z_i Z_m = \hat{P}_A(\sigma^{m-i})_{AA}E(Z_m|m \in A) \tag{5.27}$$

$$= \hat{P}_A(\sigma^{m-i})_{AA}\frac{1}{n_A}\sum_{S\cap A} Z_k \tag{5.28}$$

$$= \frac{n}{n_A}\hat{P}_A^2(\sigma^{m-i})_{AA} \tag{5.29}$$

Continuing in this way by reducing the variance of the sum of Z_i to the sum of the variances and covariances of Z_i we find that

$$\hat{V}(Z_1 + \cdots + Z_n) = \hat{V}(n\hat{P}_A) = n^2\hat{V}(\hat{P}_A) \tag{5.30}$$

$$= n\hat{V}(Z) - \hat{P}_A^2 n(n-1) + \frac{2nP_A^2}{n_A}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\sigma^{i-j})_{AA} \tag{5.31}$$

Solving for $\hat{V}(\hat{P}_A)$ gives equation 5.22.

Unfortunately, the variance estimator 5.22 is not unbiased for a couple of reasons. Firstly, $\hat{\delta_U}$ is included in each Z_i term, and will therefore affect the covariance between Z_i. This would be difficult to account for and is not included in equation 5.22. But for sufficient sample size, variance of $\hat{\delta_U}$ is generally very small, as the selection probabilities for this quantity are proportional to its size. Secondly, the transition probabilities σ are not in general known, and usually must be estimated. Although the estimation of variance is not unbiased, under most conditions it will perform quite well. Its performance is explored by simulation in the next section.

5.5 Simulation study of RDS I and RDS II

So far we have presented three estimators for RDS data, RDS I, RDS II, and RDS II/DS. In addition we have an estimate of variance for RDS II, given by equation 5.22. To gain insight into the properties of these estimators we have

performed computer simulations of RDS samples on random networks with known properties.

There are several technicalities in the simulation of RDS due to the complicated sample design. The population under study is represented by a random network, which can have a wide range of properties and can be generated by any of many algorithms that have been developed for the task. In addition, each node must have a value assigned to it for the study-variable, y_i. Secondly, random walks of specified length are executed on the random network by choosing a node uniformly at random from the network's giant component, and then randomly selecting a neighbor of the last node at each step of the random walk. These random walks are interpreted as RDS samples by keeping track of the degree of each node and the node's y_i value. In these simulations, we consider the estimation of P_A, such the y_i is the indicator function for membership in group A.

Several pieces of information are required to construct a random network:

- A specification of group sizes– that is, the size of each group A, B, ...
- A list of degree distributions for all groups in the network
- A mixing matrix $\mathscr{A}$, where the element $\left[\mathscr{A}\right]_{XY}$ specifies the fraction of all connections in the network that exist between groups X and Y.

We proceed by randomly assigning each node a degree drawn from the corresponding degree distribution. Then we randomly match connections in the network while simultaneously satisfying the constraint specified by $\mathscr{A}$.[4]

The parameter space specified by the group sizes, degree distribution, and mixing matrix is vast. In these simulations we have kept the population size fixed

[4]A more detailed description of methods for generating random networks which exhibit assortative mixing can be found in (Newman 2002).

Table 5.2: Random networks were generated with four disjoint groups, each having the size N_X and Poisson degree distribution with average degree z.

Group	N_X	z
A	1000	32
B	3000	40
C	3000	48
D	3000	56

at $N = 10,000$. The networks are divided into four groups. The variable we wish to estimate is $P_A = 0.1$, so that $N_A = 1000$. The remaining three groups are equal in size $N_X = 3000$. In addition, each group has its own degree distribution. In all cases the degree distribution is Poisson, but with different parameters controlling the average degree of the group. These variables are summarized in table 5.2.

The effects of both sample size and assortative mixing have been determined by experiment. Figure 5.5 shows the the effects of sample size on the variance of the three estimators. 5 random networks were generated, and 10,000 random walks were executed on each network. The estimators were applied to each random walk, and the empirical variance of each estimator computed. The average estimate of variance (eqn. 5.22) from these simulations is also shown in figure 5.5.

As demonstrated in section 5.3, RDS II and RDS I/DS coincide very closely. For small sample size, RDS II is more accurate than RDS I/DS, as the latter methodology relies on accurate estimation of transition probabilities to perform reliably. Both RDS II and RDS I/DS are consistently more accurate than RDS I.

In this set of simulations, the variance estimator shows slight but consistent bias

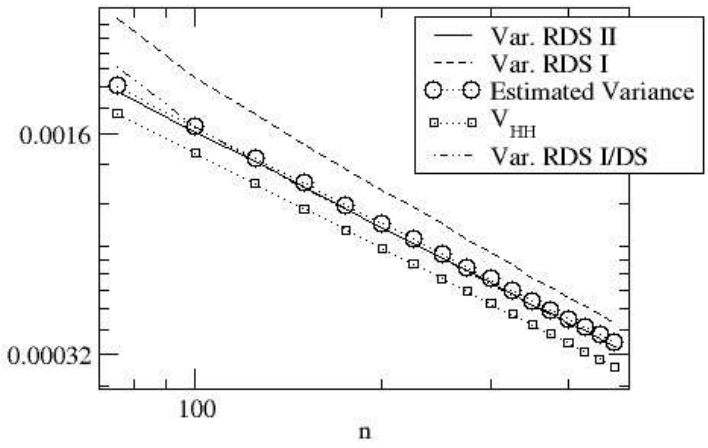


Figure 5.2: Variance of three RDS estimators and mean estimated variance, based on 50,000 simulations as described in the text. Sample size is varied from 75 to 500. The data are plotted with log-log axes.

in over-estimating the actual variance. The coverage probability of the variance estimator for 90% confidence intervals is 91.03%.

A different scenario is presented in figure 5.5. Recall that σ_{AA} is the probability that someone of type A will recruit again someone of type A. σ_{AA} actually represents an aspect of network topology: it is the proportion of connections from nodes of type A that go to other nodes of type A. In these simulations we have varied this parameter from $\sigma_{AA} = 0.069$, which corresponds to essentially no assortative mixing, to $\sigma_{AA} = 0.57$ which represents a very strong preference for nodes of type A to connect to one another at the expense of connections to nodes of other types. In all simulations, the sample size was $n = 500$.

The effect of increasing assortative mixing is the exponential increase in the variance of the estimator. In terms of MCMC sampling, this corresponds to increased sample size required for the sample to reach equilibrium.

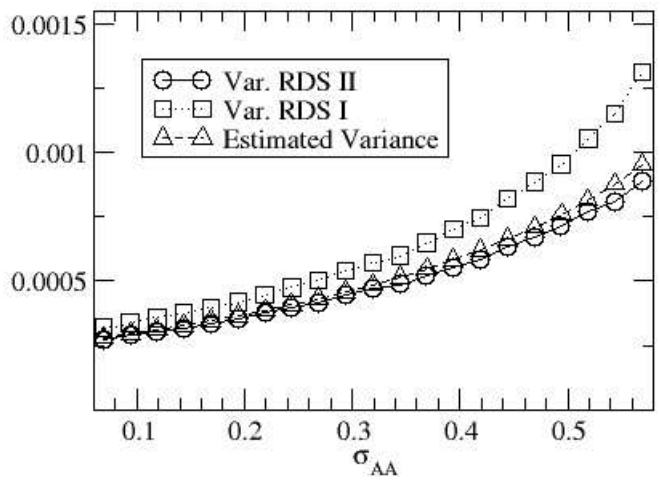


Figure 5.3: Variance of RDS II and RDS I, alongside the estimated variance for RDS II based on 50,000 simulations with sample size 500 as described in the text. The mixing parameter σ_{AA} is varied from 0.069 to 0.57.

The average coverage probability for this set of simulations is 89.997% for 90% confidence intervals. The estimated variance correctly tracks the exponential trend. The naive estimate of variance (not shown) which does not account for assortative mixing (eqn. 5.20) grossly underestimates the actual variance.

Finally, it provides some useful perspective to compare the estimators with real data. Table 5.3 shows RDS I,RDS I/DS, and RDS II, as applied to data from a study of 264 New York city jazz musicians [7]. Various categorical and continuous variables are estimated. Note that for dichotomous variables such as *Gender* and *Union membership*, RDS I and RDS I/DS give identical estimates.

In particular, note that although the variance of RDS II and RDS I/DS are generally very close, individual estimates can diverge appreciably when not demographically adjusting the RDS II estimator.

Table 5.3: RDS I, RDS I/DS, and RDS II are compared for a real data set. The data come from a survey of 264 New York city jazz musicians [7].

Estimator	Gender(Male)	Race(White)	Race(Black)	Union Membership	Age (mean)
RDS I	76.2%	53.8%	35.0%	25.1%	-
RDS I/DS	76.2%	53.2%	35.9%	25.1%	-
RDS II	72.0%	55.7%	32.8%	23.8%	42.97
Sample (Naive estimate)	73.7%	54.8%	32.8%	39.9%	45.46

5.6 Discussion

This article has further developed RDS estimation theory. A new estimator for RDS data has been presented (RDS II) which offers superior precision to prior methodology (RDS I), with the advantage of increased simplicity, analytical tractability, and analytical variance estimation. The new estimator also allows for the estimation of continuous as opposed to categorical variables. The classical RDS estimator requires quite a bit of custom code in order to derive the recruitment matrix and solve the system of linear equations (see section 5.3), and is restricted to the estimation of categorical variables[5].

There are multiple issues that still need to be addressed. The theory developed here relied on the sampling-with-replacement assumption. Biases which may be introduced due to sampling without-replacement are poorly understood. When individuals are eliminated from the pool of potential recruits, not only can they not be re-selected into the sample, but all avenues for recruitment that pass through them are also eliminated. If the average degree and population size are small, this can have unpredictable effects on the selection probabilities for everyone in the population.

[5]See *http://www.respondentdrivensampling.org* for downloads of RDS software for computing the classical RDS estimator.

In RDS samples, it is usually the case that respondents are allowed to recruit more than one person into the study. It is possible for this to introduce biases into the sample, for example if the number of recruits is correlated with the study variable or degree, however these biases remain poorly understood.

The variance estimator presented here uses the known mixing properties of the population. In general, the mixing matrix will not be known, and will have to be estimated. Furthermore, it is possible for there to exist higher-order correlations between sample elements than is presented in the mixing matrix, so that the estimated covariance between sample units may actually be biased. Such problems are inevitable whenever sampling from a network with unknown composition and structure. Certainly more could be done in improving this estimate of variance, though our simulations indicate that for most applications it should perform well.

An important problem not confronted here is how to fit models such as linear and logistic regressions to RDS data. Model-fitting should incorporate sample weights, as well as information about correlation between sample units.

The refinements to RDS theory outlined in this article should prove useful in exploring these problems, and as RDS is applied with greater frequency around the world, should find wide application.

REFERENCES

[1] W.G. Cochran. *Sampling Techniques*. Wiley, New York, 1977.

[2] P. Erdős and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[3] M.H. Hansen and W.N. Hurwitz. On the theory of sampling from finite populations. *Ann. of Math. Stat.*, 14:333, 1943.

[4] W.K. Hastings. Monte carlo sampling methods using markov chains and their their applications. *Biometrika*, 57:97–109, 1970.

[5] D. Heckathorn. Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44:174–199, 1997.

[6] D. Heckathorn. Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49:11–34, 2002.

[7] D. Heckathorn and J. Jeffri. *Changing the Beat: A Study of the Worklife of Jazz Musicians*, volume III of *Survey Results by the Research Center for Arts and Culture*, chapter Social Networks of Jazz Musicians. National Endowment for the Arts Research Division, Washington D.C., 2003.

[8] S.S. Lang. New sampling method to track hiv-risk behavior. *Medical News Today*, November 2004.

[9] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. The monte carlo method. *J. Chem. Phys.*, 21:1087, 1953.

[10] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, 2001.

[11] J. Ramirez-Valles, D. Heckathorn, R. Vázquez, R.M. Diaz, and R.T. Campbell. From networks to populations:the development and applications of respondent-driven sampling among idus and latino gay men. *AIDS and Behavior*, 9:1–16, 2005.

[12] M. Salganik and D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193–240, 2004.

[13] M.G. Sirken. A short history of network sampling. In *Proc. of the Survey and Research Methods Section*. ASA, 1998.

[14] S. Sudman and G. Kalton. New developments in the sampling of special populations. *Annual Review of Sociology*, 12:401–429, 1986.

[15] D.J. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, 1999.